NET Institute*

www.NETinst.org

Working Paper #11-13

October 2011

An Equilibrium Model of User Generated Content

Dae-Yong Ahn
Chung-Ang University

Jason A. Duan
University of Texas at Austin

Carl F. Mela
Duke University

* The Networks, Electronic Commerce, and Telecommunications ("NET") Institute, http://www.NETinst.org, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, "virtual networks" comprised of computers that share the same technical standard or operating system, and on network issues in general.

An Equilibrium Model of User Generated Content

Dae-Yong Ahn* Jason A. Duan† Carl F. Mela‡§

December 8, 2011

*Assistant Professor, College of Business and Economics, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 156-776, Korea; email: daeyongahn@cau.ac.kr; phone: 82 2 820 5944.

†Assistant Professor, McCombs School of Business, University of Texas at Austin, One University Station B6700, Austin, Texas 78712; email: duanj@mccombs.utexas.edu; phone: 512 232 8323.

‡T. Austin Finch Foundation Professor of Business Administration, Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, North Carolina, 27708; email: mela@duke.edu; phone: 919 660 7767.

§The authors would like to thank Joel Huber, Wagner Kamakura, Vineet Kumar, Oded Netzer, and seminar participants at the University of Chicago, University of Texas, and the 2011 QME Conference for their comments. The authors also thank the NET Institute (www.NETinst.org) for partial financial support of the project.

Abstract: An Equilibrium Model of User Generated Content

This paper considers the joint creation and consumption of content on user generated content platforms (e.g., reviews or articles, chat, videos, etc.). On these platforms, users' utilities depend upon the participation of others; hence, users' expectations regarding the participation of others on the site becomes germane to their own involvement levels. Yet these beliefs are often assumed to be fixed. Accordingly, we develop a dynamic rational expectations equilibrium model of joint consumption and generation of information. We estimate the model on a novel data set from a large Internet forum site and use the model to offer recommendations regarding site strategy. Results indicate that beliefs play a major role in UGC, ignoring these beliefs leads to erroneous inferences about consumer behavior, and that these beliefs have an important implications for the marketing strategy of UGC sites.

We find that user and site generated content can be either strategic complements or substitutes depending on whether the competition for existing readers exceeds the potential to attract new ones. In our data, the competitive effect substantially dilutes the market expansion effect of site generated content. Likewise, past and current content can also be either strategic substitutes or complements. Results indicate more durable content increases overall site participation, suggesting that the site should invest in making past information easier to find (via better search or page design). Third, because content consumption and generation interact, it is unclear which factor dominates in network growth. We find that decreasing content consumption costs (perhaps by changing site design or via search tools) enhances site engagement more than decreasing content generating costs. Overall, enhancing content durability and reducing content consumption cost appear to be the most effective strategies for increasing site visitation.

1 Introduction

By dramatically lowering the cost of content dissemination and consumption, online communication platforms have engendered a rapid proliferation in global user engagement. Evidence is afforded by a recent ranking done by Google's Ad Planner, listing several user sites with substantial user generated content among the top 20 most trafficked web sites (`Youtube.com`, `Wikipedia.com`, `Mozilla.com`, `Wordpress.com`, `Ask.com`, `Amazon.com` and `Taobao.com`).[1] Coincident with this increase, advertisers are spending more of their budget on social media and user generated content sites (UGC), exceeding $2BB annually, or more than 8% of firms online advertising expenditures (eMarketer 2010).

UGC platforms rely upon two behaviors; consuming content (e.g., listening or reading) and generating content (e.g., discussing or writing). Content consumption generates utility via the pleasure of reading or the utility of information. Content generation, like posting a review, yields utility from the reputational effect of being influential, knowledgeable or popular, suggesting utility increases as more of their content is consumed (Bughin (2007); Hennig-Thurau et al. (2004); Nardi et al. (2004); Nov (2007)).[2] Hence, the utility of content increases as the number of persons consuming the content (for example, reading a review) increases. Accordingly, the content generation decision is predicated on beliefs about the number of other people consuming and generating content. As such, users' beliefs about others participation on the platform are central to the problem content generation and consumption. In spite of this few, if any papers, explicitly consider the role of these beliefs on the growth of UGC networks.

We address this gap by capturing the evolution in beliefs about future consumption and generation of content, that is we allow these beliefs about the site participation of others to be endogenous. In the process, we develop a dynamic rational expectations equilibrium model of user generated content and consumption in the context of heterogeneous users. In this equilibrium, users on a UGC platform reason that the aggregate growth in the network should be consistent with sum of decisions made by all individuals who are users of it (Lee

[1] http://www.google.com/adplanner/static/top1000/

[2] In this paper we use content and posting interchangeably in which case posting implies the posting of user generated content.

and Wolpin 2006; Krusell and Smith 1998). This rational expectations equilibrium forms the basis of a joint dynamic and structural model of consumption and generation of content wherein reading is posited to increase with content availability and content availability is posited to increase with readership. Owing to its structural orientation, this approach enables us to address of number of questions of interest to UGC platforms:

1. Site Generated Content. To increase consumers' utility of consumption, platforms can provision more site generated content; for example, a site with user forums could actively participate via additional content. However, the problem of managing site generated content is challenging. On the one hand, increased site content attracts more users because of the increased availability of information. In this instance, site content is a strategic complement to user content. On the other hand, site generated content can dissuade users from posting content because site and user content are substitutes. In this instance, it is a strategic substitute. The optimal site generated content, therefore, becomes a question of the relative magnitude of these various effects. In our context, site and user content are strategic complements at low levels of site content, but become substitutes as the site content crowds the user content. The optimal 12% increase in site content would increase user traffic by 2.2%.

2. Content Durability. An analogous argument holds for the durability of user content (e.g., the ease of finding past content or its relevance).The more durable the content, the greater the potential site content available to readers thereby making the site more attractive to readers. For example, a searchable archive of past content makes older content more accessible and increases the likelihood of a reader finding relevant information. In this sense, current and past content can be strategic complements. However, the increased availability of past content also competes for reader attention with content in the current period. As a result, more durable content increases the competition for readers both from others and one's own past content. In this sense, past and current content can also be strategic substitutes. We show both effects obtain, though doubling the expected lifetime of content (from 1 to 2 weeks) increases user traffic by 9.1% and user generated content by 22.8%. Overall, this appears to be a

particularly effective strategy for the site.

3. Content Generation and Consumption Costs. Likewise, sites can lower user participation costs via site design changes, frequent participation programs or other incentives to increase engagement. In this instance, a site might be concerned about whether to weigh content generation or content consumption more heavily in designing a customer engagement program. How to weight consumption vs. generation is incumbent upon the indirect network effects of one behavior on the other. Even if the direct effect of posting content is small, for example, the indirect effect of posting content on reading can be large. We find the arc elasticity of content generation costs to be 1.89 for content generation and 0.38 for visits while the arc elasticity of content consumption costs is 0.79 for content generation and 1.13 for visits. Thus, if the cost of improving the content generation experience is less than two times the cost of consumption, the firm should emphasize an enhancement of the generation experience.

4. Beliefs About Future Consumption and Content. Owing to the dynamics in beliefs, early user trials can have a profound effect upon whether the network grows or implodes. Without sufficient reading mass, content generators might believe there is little value in creating content, thereby leading to a downward "death spiral" for the hosting platform. Related, the concept of self-fulfilling prophecies are germane in the context of social engagement, because the beliefs that others will enter the site can induce a herding behavior towards using the site. A site could create these prophecies, for example, by advertising its intention to increase overall participation. We find that changing user beliefs about future content will have little effect on the site participation in our data and therefore the equilibria are quite stable. However, this is not universally true as the theoretical model suggests that expectations are material in the early stages of the network where beliefs can affect whether the network tips or implodes.

Of note, policy experiments and comparative statics are profoundly affected when beliefs about the participation of others is not allowed to evolve with changes in the system as is common in descriptive research. For example, we find the effect of decrease in the costs of reading content on site visitations is underestimated by 26% when beliefs are not endogenized

because the potential for more readers to attract new content is neglected, leading to a 44% underestimation of content relative to the case when beliefs are exogenous.. Hence, the rational expectations equilibrium approach we use is critical when assessing how user generated content is affected by firm strategy and changes in the environment.

In sum, by integrating beliefs regarding the effect of others consumption and generation of content on one's own content decisions with a rational expectations equilibrium, our key contribution is to develop a model that enables us to explore the growth of UGC network. Though our approach is quite general and applies to many content generation and consumption contexts ranging from chat rooms to journal publications to video sharing sites (where users post and consume content), we estimate this model using a proprietary data from a web site where users generate and consume content in the form of reviews and forum postings.

In the next section, we elaborate upon how our model of user engagement differs from prior work on social networking in general, and user generated content in particular. We then discuss our data and context and use this information to construct our model. Then we explore some of the theoretical properties of our model, discuss identification, detail our results and conduct policy simulations regarding the effect of site generated content on reading and user generated content.

2 Literature Review

Our work is related to the nascent but growing empirical literature in marketing on social networking and interaction (e.g., Ansari et al. (2011), Stephen and Toubia 2010, Bulte 2007, Hartmann 2010, Nair et al. 2010, Katona et al. (2011), and Iyengar et al. 2010). Our work deviates from the social networking literature inasmuch as we consider user sites with large numbers of agents such that any single agent's participation is not likely to have a sizable effect on aggregate content consumption or generation. To exemplify this point, consider a user who posts a review on a movie site or Amazon or a video on YouTube; this agent might focus more upon the sizable number of interested viewers consuming their content than any given viewer who consumed it. In this regard, our work is analogous to the rational expectations equilibrium literature in labor economics wherein persons do not believe that their own participation in the labor force affects wages, but rather that the

aggregate participation of agents like them will. In such instances, it becomes feasible to model the dynamic social engagement choices of agents in a structural fashion because we do not need to condition on the behavior of all other individual agents (e.g., Hartmann 2010), but only the aggregate states such as the total number of posts or reads.

Likewise, our research is related to the burgeoning literature on user generated content (Albuquerque et al. 2010, Chevalier and Mayzlin 2006, Dellarocas 2006, Duan et al. 2008, Hofstetter et al. 2010, Ghose and Han 2010, Zhang and Sarvary (2011), and Zhang et al. 2011) that considers the joint behavior of content consumption and generation.[3] Our research extends this work by developing a dynamic structural model of UGC; specifically, our model allows users beliefs about site engagement to evolve with the state of the network. This is material because changes in beliefs regarding the number of users contributing, for example, can affect whether agents visit a site, consume, or write. If interventions change these beliefs, it stands to reason that the behaviors of the agents will change. It is therefore desirable for any policy intervention to accommodate potential changes in beliefs. Moreover, given that user generated content, like advertising, decays in efficacy over time and that those who post develop expectations about the likelihood their content is read in the future, there is considerable potential for dynamic behavior to be evidenced in the context of UGC.

As a dynamic structural model, our work is similar to Huang et al. (2011), who consider the blogging behavior of the employees of an IT firm. An important point of difference is that we use an "approximate aggregation" rational expectations equilibrium framework (Lee and Wolpin 2006; Krusell and Smith 1998) to link individual behavior to aggregate state transitions (such as total posting and reading). In contrast, Huang et al. (2011) model users' behaviors independently of how others at the site react. Because of the approximate aggregation approach, the equilibrium of our model can accommodate a large number of heterogeneous agents (more than 100,000 forum users in our dataset). The aggregate state transitions across all the users can vary with changes in the primitives of the system, yielding a structural interpretation of the social engagement problem. To our knowledge, ours is the

[3]Ghose and Han (2011) consider a dynamic structural model of mobile phone content usage based on consumer learning; they do not jointly model the dynamics in consumption and content generation. Our work is also complementary to theirs inasmuch as the dynamics in our model reflect expectations about future readership for posts rather than uncertainty in the usage experience.

first application of approximate aggregation models to marketing, and this approach can generalize to other marketing contexts involving large numbers of heterogeneous agents who condition on the aggregation of others behaviors rather than the behavior of any other particular agent.

In sum, our contribution is to develop a dynamic structural model of content generation and consumption for a large number of users and use this model to evaluate indirectly network effect in a dynamic setting and draw implications regarding how the site who hosts these interactions should manage their content.

3 Model

3.1 Model Overview

Figure 1 outlines the modeling context. Users consume content generated by others owing to their interest in information. An increase in content can lead to an increase in use because users are more likely to find information of interest (Stigler (1961)). Hence, we consider a model of information search for content of heterogeneous quality. We discuss and model this indirect network effect in Section 3.2.

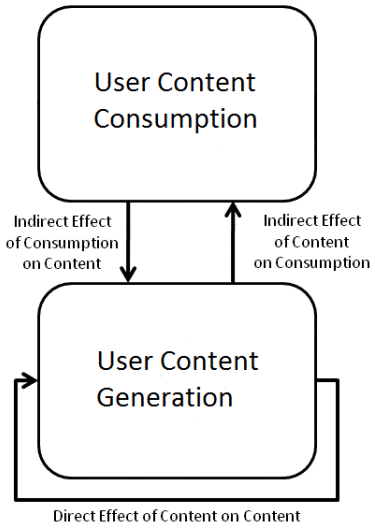

Figure 1: Model Overview

An increase in content consumption can lead to an increase in content because those who

post content presumably do so because they are motivated to have their posts read (Bughin (2007); Hennig-Thurau et al. (2004); Nardi et al. (2004); Nov (2007)) and we model this process in Section 3.3. There is also a potential direct network effect of content on content; as more content appears competition for readers increases; this problem is also considered in Section 3.3.

Last, it is theoretically possible for an increase in users to lead to more use via word of mouth (a direct effect of consumption on consumption). Our application considers a relatively short time horizon. As such, we assume these word of mouth effects to be negligible. Moreover, users are geographically dispersed and typically know each other only by their user ids. Therefore they primarily contact other users on the site via an intermediating effect of posts, but this interaction occurs only through the content they generate and is thus is an indirect network effect, not a direct effect. Moreover, given the supply of content is limitless, we expect no competitive effect between readers for content, thereby mitigating another source of potential direct network effects for consumption.

Both the decisions to generate and consume content are incumbent upon the decision to visit the site on a given day. To the extent the posting or reading utility exceeds that of outside options, users visit the site. We discuss this process in Section 3.4.

In sum, we consider M users' decisions ($i = 1, \ldots, M$) to visit a content sharing website on occasion t ($t = 1, \ldots, T$) and conditioned on that visitation decision, $n_{it} \in \{0, 1\}$, how much content to consume, r_{it}, and how much content to generate, a_{it}. Users choose each of these three actions $\{n_{it}, r_{it}, a_{it}\}$to maximize their utility conditioned on their beliefs regarding overall participation of others in the network.

3.2 Reading

3.2.1 Reading Utility

We presume that individuals read user generated content so long as its marginal benefit exceeds its marginal cost $c(r_i)$. This tradeoff between the cost of reading and gaining utility from information of interest determines the optimal number of posts that users read. As the number of posts increases, the likelihood that a user search results in items of interest increases, thereby leading to an increase in posts read. Following Stigler (1961)'s model of

information search under a uniform distribution of quality of K_t posts over the range from L (lower bound on post quality) to U (upper bound on post quality), Appendix A derives reader i's utility for reading r posts as

$$u\left(r_i\right) = \alpha_1 r_{it} - \frac{\alpha_2 r_{it}^2}{2K_t}, \tag{1}$$

where α_1 indicates the upper limit, U, on perceived content quality and $\alpha_2 \equiv U - L$ indicate the range of content quality, K_t is the content stock defined in Section 3.2.2 below. Equation 1 implies that the utility of reading evidences decreasing marginal returns in the amount of UGC and that an increase in UGC increases both the utility and the marginal utility of reading. This result follows intuitively from a greater likelihood of finding content of interest.

3.2.2 Posting Stock

The utility of reading in equation (1) is incumbent upon the stock of posts generated by users. Following the advertising literature, we assume that posted information follows a geometric decay over time (Clarke (1976), Dubé et al. (2005)). This geometric decay can be justified via probabilistic awareness. A fresh posting in period t will usually be near the top of a forum, so the probability it will be noticed by a reader is close to one. Postings from the preceding period have a lower probability of being noticed ($\rho < 1$) as they age. Following this logic, posts in period $t-k$ have a ρ^k chance of being noticed. Thus, at any given site visit, individual i will only notice $\tilde{A}_{t-k}$ of the $A_{t-k} = \sum_{i=1}^{M} a_{it-k}$ postings in period $t-k$ where $\tilde{A}_{t-k}$ follows a binomial distribution $\tilde{A}_{t-k} \sim Bin\left(A_{t-k}, \rho^k\right)$. Summing the noticed posts across periods leads to an aggregate stock formulation of form:

$$K_t = E\left(\sum_{\tau=1}^{t} \tilde{A}_\tau\right) = \sum_{\tau=1}^{t} \rho^{t-\tau} A_\tau = \rho K_{t-1} + A_t$$

where $\rho < 1$ is the discount rate and A_τ is the number of postings in period t. Likewise, the individual-level stock of postings is given by

$$k_{it} = \sum_{t=1}^{\tau} \rho^{\tau-t} a_{i\tau} = \rho k_{i,t-1} + a_{it}. \tag{2}$$

The geometric decay argument can analogously be extended to site generated content leading to the total stock of information being given by:

$$K_t = K_t^U + K_t^S, \tag{3}$$

where K_t^U denotes user generated content and K_t^S is an analogously constructed variable that captures the stock of site generated content.

3.2.3 Reading Costs

Next we consider the cost of reading. Following Yao and Mela (2008) and others, we assume the cost has a quadratic form that reflects an increasing scarcity of time or attention as more items are read,

$$c\left(r_{it}\right) = \kappa_1 r_{it} + \kappa_2 \frac{r_{it}^2}{2}, \tag{4}$$

where κ_1 and κ_2 are positive implying increasing marginal cost and $\kappa_2 > 0$ means an increasing marginal cost of reading. Hence, the total payoff of reading is

$$u\left(r_{it}\right) - c\left(r_{it}\right) = \left(\alpha_1 - \kappa_1\right) r_{it} - \left[\frac{\alpha_2}{K_t} + \kappa_2\right] \frac{r_{it}^2}{2}. \tag{5}$$

Given this utility, the optimal reading r_{it}^* is solved by the first order condition

$$\begin{aligned} \frac{d}{dr_{it}} \left[u\left(r_{it}\right) - c\left(r_{it}\right)\right]|_{r_i = r_i^*} = \left(\alpha_1 - \kappa_1\right) - \left[\frac{\alpha_2}{K_t} + \kappa_2\right] r_{it}^* = 0 \Rightarrow \\ r_{it}^* = \frac{\alpha_1 - \kappa_1}{\alpha_2 / K_t + \kappa_2} \end{aligned} \tag{6}$$

3.2.4 Heterogeneity and Seasonality Effects

We model reading heterogeneity in the cost function as follows: a random effect for unobserved time-invariant heterogeneity ζ_i and an individual and per-period random shock ν_{it}, which allows for unobserved situational error.

We accommodate seasonality by allowing κ_1 in (4) to vary overtime and re-define it as κ_{1it}. This variation allows for seasonal effects such as work week effects where broadband access is often higher.

The seasonal effect κ_{1it} is also indexed by i because it can be heterogeneous across user. The coefficient of quadratic term in the reading cost function κ_2 can also be heterogeneous, so we let it be κ_{2i}.[4]

Accordingly, the payoff of reading in equation (5) can be rewritten as

$$u\left(r_{it}\right) - c\left(r_{it}\right) = \nu_{it}\left(\alpha_1 - \kappa_{1it} - \zeta_i\right) r_{it} - \left[\frac{\alpha_2}{K_t} + \kappa_{2i}\right] \frac{r_{it}^2}{2}, \tag{7}$$

[4]Equation (6) indicates that only $\alpha_1 - \kappa_1$ determines r^* in the numerator. Hence, a constraint has to be imposed on κ_{1it} to identify α_1 in the empirical analysis.

where ν_{it} is the random shock in the amount of reading across time and is assumed iid across individuals and time periods. As users typically read in large quantities, we treat r_{it} as a continuous variable and hence assume ν_{it} to be continuous. [5] We further assume ν_{it} is independent of ζ_i, κ_{1it} and κ_{2i} and all of them are known to the users but not the researcher.[6] Because ν_{it} is realized after a user's decision on whether she visits the forum, we can further assume v_{it} independent of the site visitation decision and $E(\nu_{it}|n_{it}) = 1$.[7] .

To model the heterogeneity in ζ_i, κ_{1it} and κ_{2i}, we use a finite mixture model by letting them follow a discrete distribution which can be interpreted as a finite number of latent segments of readers

$$[\zeta_i, \kappa_{1it}, \kappa_{2i}] \sim \sum_{j=1}^{J} p_j I\left(\zeta_i = \bar{\zeta}_j\right) I\left(\kappa_{1it} = \bar{\kappa}_{1jt}\right) I\left(\kappa_{2i} = \bar{\kappa}_{2j}\right),$$

where there are J latent classes and user i belongs to class j with probability p_j. We constrain $\bar{\zeta}_j$ such that $\sum_{j=1}^{J} \bar{\zeta}_j = 0$ because α_1 already subsumes a non-zero mean for readings.

With these errors, the optimal reading level in 6 can now be expressed as

$$r_{it}^* = \frac{\alpha_1 - \kappa_{1it} - \zeta_i}{\alpha_2/K_t + \kappa_{2i}} \nu_{it}. \tag{8}$$

Given that ν_{it} is realized after the user's site visitation decision, the user's decision to visit the forum at time t depends only on the expected optimal amount of reading defined by

$$E_i\left(r_{it}^* | K_t, \zeta_i\right) = \frac{\alpha_1 - \kappa_{1it} - \zeta_i}{\alpha_2/K_t + \kappa_{2i}} E\left(\nu_{it}\right) = \frac{\alpha_1 - \kappa_{1it} - \zeta_i}{\alpha_2/K_t + \kappa_{2i}}, \tag{9}$$

and the expected optimal number of postings, which will be discussed in Section 3.3.

[5] It is not imperative to impose any parametric distribution on ν_{it} though we assume ν_{it} to be exponential in Appendix E.1 to facilitate maximum likelihood estimation.

[6] The ν_{it}'s are also not correlated with K_t, though K_t is an endogenously generated variable. This is because if we have a large number M of users, then $1/M \sum_{i=1}^{M} \nu_{it} \approx E[\nu_{it}] = 1$ by the law of large numbers. We will see in the following subsections that users only use the aggregate readings $R_t = \sum_{i=1}^{M} r_{it}$ in their posting decision problem. When M is large, $R_t = \sum_{i=1}^{M} r_{it}^* \nu_{it} \to M r_{it}^*$, so the effect of v_{it}'s will cancel out.

[7] Users first decide whether they will visit the forum website at t before they make decision on the amount of reading and the number of postings. The random shock in reading and posting payoffs are both assumed to be realized after the visiting decision is made. Section 3.4 details how the site visitation decision is made based on expected maximal posting and reading payoffs.

3.3 User Generated Content

3.3.1 The Per-Period Utility of UGC

Site users derive utility from others reading their posts and this expected posting utility is incumbent upon the users beliefs their postings will be read. The expected average amount of reading per posting is used to model the reading likelihood because a user on our forum cannot observe the exact amount of reading for each of her postings (there is not a counter of "number of views" on the forum we model unlike, e.g., Youtube). This expected amount of reading per posting (y_t) is defined by

$$y_t = \frac{R_t}{K_t} = \frac{\sum_{i=1}^{M} E(n_{it} r_{it}^* | K_t, \zeta_i)}{K_t}. \tag{10}$$

Equation (10) demonstrates two competing effects aggregate UGC K_t on y_t. First, there is a primary demand effect of K_t in the numerator as the expected optimal amount of reading increases with the supply of content, K_t based on equation (9). Second, there is a competitive effect of K_t in the denominator as more postings will reduce the amount of reading per posting. Therefore, the net effect K_t on y_t can be positive or negative. In Appendix B, we show that the user's expected amount of reading per posting y_t can be closely approximated by the exactly observed amount of reading per posting under the assumption of rational expectations, when the number of users and the UGC stock K_t are both very large. Hence, in the subsequent model development, we do not distinguish between expected and observed reading rates.

Given the imputed likelihood their posts are read, the current period expected utility from generating content in period t can be written as a function of the number of the posts a user i writes and the rates with which these posts are read,

$$u(a_{it}|s_{it}) = g(k_{i,t} y_t) = g\left(\left[\rho k_{i,t-1} + a_{it}\right] y_t\right), \tag{11}$$

where $g(x)$ is a utility function with diminishing marginal return. A common choice of $g(x)$ is

$$g(x) = \frac{x^{1-\gamma}}{1-\gamma}, \gamma \in [0, \infty) \tag{12}$$

where we have $g(x) = x$ when $\gamma = 0$ and $g(x) = \log(x)$ when $\gamma = 1$.

In equation (11), s_{it} is defined as a set of state variables which we enumerate next. First, k_{it} is the posting stock of user i at period t, so $k_{it} \in s_{it}$. Second, the expected amount of reading per posting y_t defined by equation (10) is merely a function of the aggregate posting stock K_t and hence $K_t \in s_{it}$. Other state variables in the UGC posting problem will appear in the posting cost function, which will be defined in the following subsection. The actions variable a_{it} represents the number of postings user i choose at period t given her state variables k_{it} and K_t. This posting decision a_{it} is comprised of a discrete set of integer number of posting, i.e., $a_{it} \in A = \{0, 1, 2, \ldots, \bar{a}\}$, where $\bar{a}$ is a large integer representing the upper bound of postings a user can write in period t.

3.3.2 The Cost of UGC Posting

Similar to the cost of reading, the cost of writing is specified as

$$c_{it}(a_{it}|s_{it}, \varepsilon_{it}) = (\tau_{it} + \xi_i)a_{it} - \varepsilon_{it}(a_{it}), \tag{13}$$

where, the random errors in the cost function, $\varepsilon_{it}(a_{it})$, has a generalized extreme value (GEV) distribution, which will be specified below with the site-visitation model (Section 3.4) to form a nest-logit type of choice probabilities. The time-invariant component of the linear marginal cost ξ_i is heterogeneous across users and is assumed to follow a discrete distribution for a latent segment model. τ_{it} models seasonal effect such as a weekend effect. We also assume the seasonal effect τ_{it} to be idiosyncratic to different latent segments. Together with the heterogeneity in the reading cost function, we propose the following joint discrete distribution for the latent segment model:

$$[\xi_i, \tau_{it}, \zeta_i, \kappa_{2i}] \sim \sum_{j=1}^{J} p_j I\left(_i = \bar{\xi}_j\right) I\left(\tau_{it} = \bar{\tau}_{jt}\right) I\left(\zeta_i = \bar{\zeta}_j\right) I\left(\kappa_{2i} = \bar{\kappa}_{2j}\right), \tag{14}$$

where there are J latent segments and user i belongs to segment j with probability p_j. In the formula above, $\bar{\xi}_j$ is the segment-specific value of the time-invariant effect in the marginal cost of posting if user i belongs to segment j and $\bar{\tau}_{jt}$ is segment-specific seasonal effect. Finally, we define the non-random part of cost $c_{it}\left(a_{it}|s_{it}\right)$ to be

$$\bar{c}_{it}\left(a_{it}|s_{it}\right) = (\tau_{it} + \xi_i)a_{it}.$$

Because the τ_{it} and varies over time, it also enters the set of state variable s_{it}, i.e., $\tau_{it} \in s_{it}$.

3.3.3 Optimal UGC Posting Levels

Given the decay in stock, a forum user's posting decision becomes dynamic optimization problem. A user chooses the number of posts (amount of content to generate) a_{it} that maximizes the discounted expected sum of period utilities minus period costs to obtain the following value function

$$V_i\left(a_{it}|s_{it},\varepsilon_{it}\right)=\max_{a_{it},a_{i,t+1},\ldots}E\left(\sum_{\tau=t}^{\infty}u(a_{it}|s_{it})-c_{it}(a_{it}|s_{it},\varepsilon_{it})\right). \tag{15}$$

In this dynamic optimization problem, $s_{it}=\{k_{it},K_t,\tau_{it}\}$ and ε_{it} are the state variables and the number of per-period postings a_{it} is the control variable.

The value function of this optimization problem in the form of Bellman's equation is

$$V_i(s_{it},\varepsilon_{it})=\max_{a_{it}\in A}\{u(a_{it}|s_{it})-\bar{c}_{it}(a_{it}|s_{it})+\varepsilon_{it}(a_{it})+\beta E[V_i(s_{i,t+1},\varepsilon_{i,t+1})|s_{it},a_{it}]\}, \tag{16}$$

where $A\equiv\{0,1,\ldots,\bar{a}\}$ is the action space.[8]

3.4 Site Visitation

Prior to posting, a user must first decide whether to visit the UGC website and this decision is predicated upon the expected utility from consuming and generating content should the user decide to visit. Hence, the utility from visiting the site on a given occasion includes utilities from writing and expected reading is given as

$$\begin{aligned}u(n_{it}=1|s_{it}) &= \mu_1 E\max_{r_{it}}[u(r_{it})-c_{it}(r_{it})]+\\ &\quad \max_{a_{it}}[u(a_{it}|s_{it})-c_{it}(a_{it}|s_{it},\varepsilon_{it})+\beta\tilde{EV}_j(s_{it},a_{it})]+\eta\varepsilon_{it}(n_{it}=1)\end{aligned} \tag{17}$$

where μ_1 is a scale parameter such that the utility of reading can be rescaled to the utility measure of posting.[9] The indicator variable $n_{it}\in\{0,1\}$ indexes the site visitation decision.

[8]The inter-temporal substitution of posting for the dynamic optimization problem is as follows. If we treat a_{it} as a continuous variable, we can derive the Euler equation $-\left[\left(y_t\left(\rho k_{i,t-1}+a_{it}\right)\right)^{-\gamma}-\left(\tau_{it}+\xi_i\right)\right]/(\tau_{i,t+1}+\xi_i)=\beta\rho$, which shows posting one more unit of UGC will gain utility $\left[y_t\left(\rho k_{i,t-1}+a_{it}\right)\right]^{-\gamma}$ and incur cost $(\tau_t+\xi_i)$ in the current period t. This additional posting will also gain utility discounted by $\beta\rho$ in the next period $t+1$. However, if a user selects to post in $t+1$ instead of t, she will forgo the utility in t, which is $\left[y_t\left(\rho k_{i,t-1}+a_{it}\right)\right]^{-\gamma}$, and avoid cost $(\tau_{it}+\xi_i)$. The cost incurred in $t+1$ is $(\tau_{i,t+1}+\xi_i)$ instead. The optimal number of postings is achieved when the user is indifferent about whether posting an additional unit in t or $t+1$. Thus, increasing durability of UGC, ρ, tends to increase the incentive to post in the current period. However, the competitive effect from the increased postings of other users and one's own past postings constitutes indirect disincentive to post.

[9]Regarding the expected utility of reading, note that Equation (8) assumes that users who visit a site will always read at least some posts, because $r_{it}^*>0$ is always an interior optimal solution. This implies the

The contextual shock $\varepsilon_{it}(n_{it}=1)$, which represent the exogenous cost for a user to visit the site at period t and is known to the user but not the econometrician.

The corresponding utility from not using the site contains three components. First, users continue to obtain utility from those who read their posts from past visits. The decayed posts are given by $k_{it} = \rho k_{i,t-1}$, which the attendant reading rate y_t in period t. Hence, the flow utility of other users reading posts when the site is not visited is given by $g(k_{it}y_t)$. Second, μ_{0i} is a segment-specific intercept which measures opportunity utility gained outside if the time spent on the forum is used elsewhere. Third, there is again the contextual shock $\varepsilon_{it}(n_{it}=0)$. Therefore, the utility of not visiting the site is given by

$$u(n_{it}=0|s_{it}) = \mu_{0i} + g(k_{it}y_t) + \beta\tilde{EV}_i(s_{it}, n_{it}=0) + \eta\varepsilon_{it}(n_{it}=0), \tag{18}$$

where $\tilde{EV}_i(s_{it}, n_{it}=0)$ will be defined when we derive the user's site visitation and posting probabilities in the following subsection. A user chooses to visit the forum website if $u(n_{it}=1|s_{it}) > u(n_{it}=0|s_{it})$ and vice versa.

3.5 Choice Probabilities for Content Generation and Site Visitation

To derive the choice probabilities for posting and site visitation, we assume $\varepsilon_{it}(a_{it})$, $\varepsilon_{it}(n_{it}=0)$ and $\varepsilon_{it}(n_{it}=1)$ have iid Type-1 Extreme Value (Gumbel) distributions.[10]

We further define $\tilde{EV}_i(s_{it}, a_{it})$ to be the integrated value function,

$$\tilde{EV}_i(s_{it}, a_{it}) = \int_{s_{i,t+1}} \int_{\varepsilon_{i,t+1}} V_i(s_{i,t+1}, \varepsilon_{i,t+1}) p(s_{i,t+1}, \varepsilon_{i,t+1}|s_{it}, \varepsilon_{it}, a_{it}) ds_{i,t+1} d\varepsilon_{i,t+1}, \tag{19}$$

which is the fixed point of the following functional equation (Rust, 1987 and 1994) under the conditional independence assumption for ε_{it} and s_{it}

$$\tilde{EV}_i(s,a) = \int_{s'} \log\left(\sum_{a'\in A} \exp\left\{u(a'|s') - \bar{c}_{it}(a'|s') + \beta\tilde{EV}_i(s',a')\right\}\right) \cdot p(s'|s,a)ds'. \tag{20}$$

expected utility of reading is always greater than zero if a user decide to visit. This specification is consistent with the data as 99.998% of the users read postings upon entering the site.

[10] An alternative model for the random errors assuming McFadden's Generalized Extreme Value (GEV) distribution for $\varepsilon_{it}(a_{it})$ and $\varepsilon_{it}(n_{it}=0)$ and $\varepsilon_{it}(n_{it}=1)=0$ yields a nested logit model with the equivalent inclusive value function and choice probabilities subject to reparameterization. See Choi and Moon (1997) for the details of the inclusive value function for the GEV model.

Using Equation 19, the optimal posting decision is to choose a_{it} if and only if

$$u(a_{it}|s_{it}) - \bar{c}_{it}(a_{it}|s_{it}) + \varepsilon_{it}(a_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}) \geq$$
$$u(a'_{it}|s_{it}) - \bar{c}_{it}(a'_{it}|s_{it}) + \varepsilon_{it}(a'_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}), \forall a'_{it} \neq a_{it} \in A, \tag{21}$$

by which we derive the probability of writing a_{it} forum postings conditional on site visitation as

$$P(a_{it}|s_{it}, n_{it} = 1) = \frac{\exp(u(a_{it}|s_{it}) - \bar{c}_{it}(a_{it}|s_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it}))}{\sum_{a'_{it} \in A} \exp(u(a'_{it}|s_{it}) - \bar{c}_{it}(a'_{it}|s_{it}) + \beta \tilde{EV}_i(s_{it}, a'_{it}))}. \tag{22}$$

Because not visiting the forum site leads to zero postings and the same decay of posting stock as writing no posting when the user visits the site, we have

$$\tilde{EV}_i(s_{it}, n_{it} = 0) = \tilde{EV}_i(s_{it}, n_{it} = 1, a_{it} = 0). \tag{23}$$

We define the inclusive value of writing forum postings conditional on site visitation as

$$IV_{it} = \ln \sum_{a_{it} \in A} \exp(u(a_{it}|s_{it}) - \bar{c}_{it}(a_{it}|s_{it}) + \beta \tilde{EV}_i(s_{it}, a_{it})). \tag{24}$$

Based on equations (23) and (24), we derive the choice probability of visiting the site as

$$P\left(n_{it} = 1|s_{it}\right) = \tag{25}$$
$$\frac{\exp\left\{\mu_1 E \max_{r_{it}}[u(r_{it}) - c_{it}(r_{it})] + \eta IV_{it}\right\}}{\exp\left\{\mu_{0j} + \eta\left[g(k_{it}y_t) + \beta \tilde{EV}_j(s_{it}, n_{it} = 0\right]\right)\right\} + \exp\left\{\mu_1 E \max_{r_{it}}[u(r_{it}) - c_{it}(r_{it})] + \eta IV_{it}\right\}}$$

and $P(n_{it} = 1|s_{it}) = 1 - P(n_{it} = 1|s_{it})$.

Note that when we apply the latent segment model in equation (14), the integrated value function $\tilde{EV}_i(s, a)$ is the same for all the users in segment j ($j = 1, \ldots, J$). Hence, we let $\tilde{EV}_i(s, a) = \tilde{EV}_j(s, a)$ if user i is in segment j.

3.6 State Transitions

In this section, we detail the state transition, $p(s'|s)$, indicated in equation 19. First, The individual stock k_{it} evolves deterministically $k_{it} = \rho k_{i,t-1} + a_{it}$. Second, the aggregate stock of site content, K_t, consists of two parts: the stock of site-generated content, K_t^S and the aggregate user-generated content $K_t^U = \sum_{i=1}^{M} k_{it}$. Site content, K_t^S, is exogenous and evolves stochastically over time. The aggregate UGC, K_t^U, evolves deterministically given K_{t-1}^U and

$a_{it};\ i = 1, \ldots, M$. However, from the perspective of any individual user i, K_t^U given K_{t-1}^U is stochastic because she does not observe a_{it} and ε_{it} of other users. When the site has a very large number of users, we can assume any user i believes her own action a_{it} has no influence on the aggregate UGC, K_t^U. This claim is similar to the assumption of pure competition where no agents in the market assume their individual output can change the total supply. Hence, user i assumed that K_t^U evolves given K_{t-1}^U and K_t^S, but independent of her own action a_{it}. If we impose a rational expectations constraint, then user i's belief about the state transition for K_t^U must coincide with the actual behavior by users on the site. This will be discussed in detail next in Section 3.7 below. Finally, the random shocks, ε_{it}, are assumed to be i.i.d. over time and across individuals and independent of s_{it}.

3.7 Rational Expectations Equilibrium and Approximate Aggregation

Aggregate content, $K_t = K_t^U + K_t^S$, is the sum of individual users' actions plus the exogenous site generated content.Rational expectations require that the beliefs about the K_t be consistent with its actual transitions, which reflect the sum of all individuals' posting behaviors. This observation becomes critically important in policy simulations because there is no reason to presume the evolution of K_t is invariant to a change in policy that might affect users' participation levels.

Using an approximate aggregation approach to rational expectations equilibrium pioneered by Krusell and Smith (1998), we first formulate agent's beliefs on how the aggregate state variable K_t evolve over time as follows

$$K_t^S = \omega_0^S + \omega_1^S K_{t-1}^S + \nu_t^S, \tag{26}$$

$$K_t = \omega_{0t}^U + \omega_1^U K_{t-1}, \tag{27}$$

$$\text{where } K_t = K_t^U + K_t^S.$$

The parameters $\omega_0^S,\ \omega_1^S$ relate to exogenous evolution of site-generated content K_t^S, which we set as one of primitives of the model. The amount of SGC, or K_t^S, is determined by some unobserved exogenous shock ν_t^S. Viewed in this light, equation 26 approximates the site's content generation policy (note that in the empirical data, these site generated posts

are negligible; hence this approximate policy function is used only in the counterfactual analysis).

The parameters $\omega_{0t}^U,\ \omega_1^U$[11] for the stock of the aggregate user generated content are determined by the rational expectations equilibrium. We posit the order of the lag in the state transitions to be consistent with the primitives in the consumer model to help ensure that the approximate beliefs regarding the aggregate state transitions are consistent with the Markovian structure in the underlying individual posting model.[12]

Our model also assumes individual users approximate the average amount of reading per posting as a function of K_t with

$$y_t = \omega_{0t}^y + \omega_1^y K_t. \tag{28}$$

Equation (28) approximates equation (10) which does not have a closed form for the function y_t of K_t. The parameters $\omega_{0t}^y,\ \omega_1^y$ are also determined by the rational expectations equilibrium.

We use the approximate aggregation approach similar to Krusell and Smith (1998) and Lee and Wolpin (2006) in lieu of other rational expectations approaches in marketing (e.g., Dubé et al. (2010)) because of user heterogeneity in posting stock. Though K_t is deterministic given the actions of all individuals and the solution of the implicit function

$$K_t = K_t^S + \rho\left(K_{t-1} - K_{t-1}^S\right) + \sum_{i=1}^{M} a_i\left(k_{it}, K_t, \tau_{it}, \varepsilon_{it}\right), \tag{29}$$

using equation 29 directly to compute users' rational expectations requires us to assume all users know all other users' policy functions $a_i\left(k_{it}, K_t, \tau_{it}, \varepsilon_{it}\right)$ as well as the distribution of their individual-level posting stock k_{it}. Complete knowledge of the behavior of many

[11] ω_{0t}^U is indexed by time t to incorporate the fixed effect for weekend in our empirical model. The same applies to ω_{0t}^y below.

[12] Note that the order of the state transition equations can not be higher than the order of the individual level model, else the individual level model would fail to account for consumer's beliefs about these higher order states. Here we assume individuals only use one lagged K_t to predict K_{t+1} and hence it implies an AR(1) model for K_t. Individuals may use more than one lagged K_t to predict K_{t+1}. For example, should users consider an AR(q) model $K_{t+1} = \omega_{0,t+1}^U + \omega_1^U K_t + \cdots + \omega_q^U K_{t-q+1}$, then $K_{t-1}, \ldots, K_{t-q+1}$ will also be in the set of state variables in individual i's dynamic optimization problem. As more state variables can eventually cause the curse of dimensionality, the most parsimonious state transition model for K_t is desirable. Indeed, as a robustness check for our data analysis in Section 7.2, we find the second lag coefficient ω_2^U to be nonsignificant (p-value = 0.73). Durbin-Watson test for the residuals of the AR(1) model $K_t = \omega_{0t}^U + \omega_1^U K_{t-1}$ has the p-value equal to 0.63, which cannot reject the null hypothesis that the autocorrelation of the residuals is 0.

thousands of others is an unrealistic assumption which imposes a large informational burden on every individual user. In addition, this assumption places the distribution of k_{it} in every user's set of state variables. Because the distribution of k_{it} is high-dimensional, the "curse of dimensionality" renders the dynamic programming problem intractable. As a result, the approximate aggregation (Krusell and Smith, 1998 and Lee and Wolpin, 2006) only requires that K_t and y_t computed from equations (27) and (28) respectively in the individual optimization problem coincide with K_t and y_t computed from the exact aggregation. Krusell and Smith (1998) show that using the state transition rule such as in (27) can still generate a stationary distribution of states like k_{it} instead of a degenerate k_t for every agent (which we further confirm by simulation in Section 4.2.1).

The approximate aggregation approach requires that we ensure that the aggregate state transitions are consistent with the individual behaviors that underpin it. Using an initial guess for the parameters ω_{0t}^U, ω_1^U and ω_{0t}^y, ω_1^y, we compute individual behaviors n_{it}, a_{it} and r_{it}^*. Aggregating across persons, we recompute K_t and y_t and recompute individual behaviors, iterating back and forth between the individual-level and aggregate models until convergence. Appendix C details the algorithm used to compute a rational expectations equilibrium. The parameters ω_{0t}^U, ω_1^U and ω_{0t}^y, ω_1^y are re-estimated in every step of the iterations to find the fixed point of the rational expectations equilibrium.[13] This process ensures that the users' beliefs about the aggregate state transitions are consistent with the underpinning individual behaviors. In sum, the use of approximate aggregation enables us to accommodate heterogeneity in a rational expectations equilibrium model.

4 Theoretical Implications

In this section, we explore some of the theoretical properties of our model. Specifically, we assess i) convergence to the defined rational expectations equilibrium in Section 3.7 and ii) how the model's parameters and exogenous states influence the network's user content and readings in equilibrium. This analysis studies the role of initial content on network size, how site postings affect site traffic, and the effect of content stock decay on content generation.

[13] In estimation, the aggregate states are observed (reflecting the current equilibrium), so no iteration to to re-estimate ω_0^U, ω_1^U and ω_0^y, ω_1^y is necessary.

4.1 Simulation Initialization

We consider two segments in our 100 period simulation, each of whom have the same cost of reading buy vary in their posting costs and size (one segment is smaller and has lower posting costs, consistent with the notion that a small number of users predominate the number of posts. In Appendix D we detail the specific parameters of our simulation.

4.2 Simulation Results

4.2.1 Initial Individual Stock

We select two different sets of values for the initial endowment of individual posting stocks. The first set of values has the posting stock equal to 3 for any individual in Segment 1 and 0.1 for Segment 2; the second has 8 for Segment 1 and 0.1 for Segment 2. Neither of these 2 sets of initial values are considered extremely high or low, so we expect they converge to the same equilibrium.

In Figure 2, we plot the equilibrium path of the aggregate user generated postings (UGC) after the rational expectations equilibrium is achieved. We can see that the first set of initial values (solid curve) and the second (dashed curve) converge to the same steady-state aggregate UGC with small random variations. We also find the same equilibrium parameter values in the equations (27) and (28). The UGC reaches the steady state after only about 10 periods.

Based on the theoretical model in Section 3.3, we expect that that not only the aggregate UGC converges (shown in Figure 2), but the distribution of individual posting stocks would be constant in the steady state as well. Figure 3 shows the distribution (histogram) of individual posting stocks of the two segments of site users in period 50 and 100, when the UGC has already reached the steady state. These histogram plots confirm our conjecture that these distributions are indeed invariant over time.

4.2.2 Degenerate Equilibrium

One potential equilibrium of our model is that all individual postings, amount of reading and site visiting are zero. That is, the network will never expand unless some shock or intervention enables the network to tip from a non-zero state. For example, extremely low

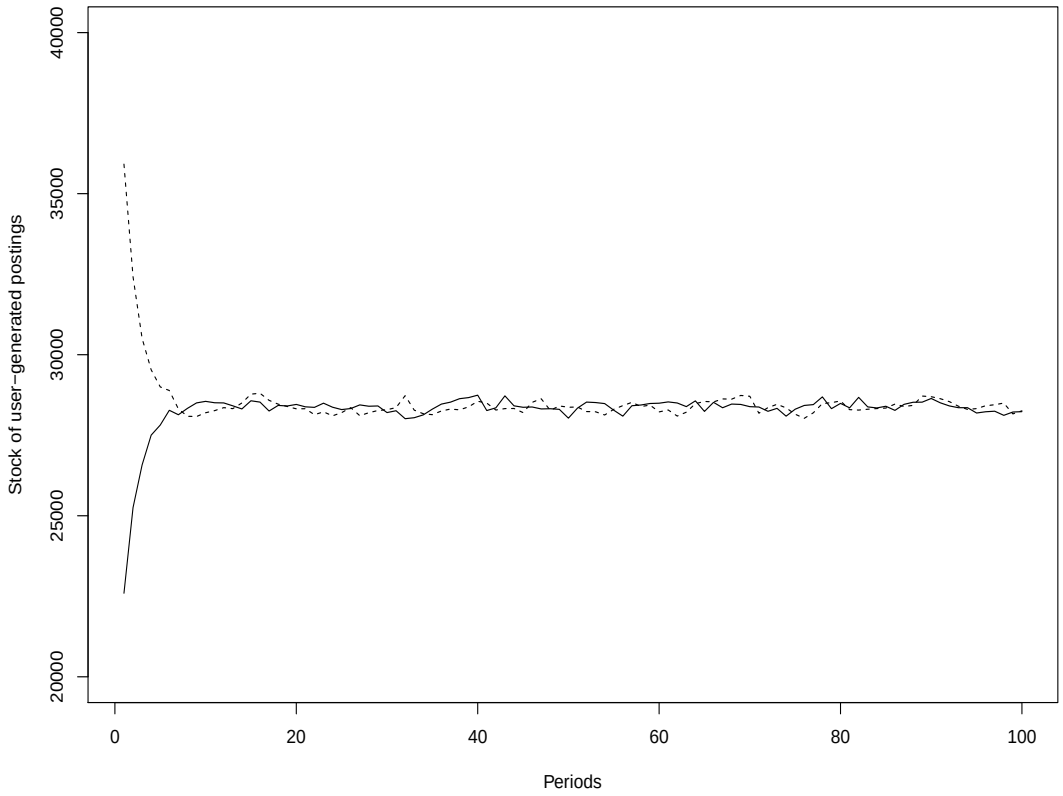

Figure 2: Convergence of aggregate user-generated posting stock (UGC) to the steady state from 2 different starting values.

user-generated posting stock can cause the low reading and site visiting rate, which can in turn cause lower posting activity and even lower posting stock. In order to test this conjecture, we select a set of very low initial values for posting stocks: 0.1 for both Segments 1 and 2. The dashed curve in Figure 4 demonstrates the result of this simulation which converges to the trivial equilibrium, which implies low user activity can eventually cause the forum to collapse.

4.2.3 Site-generated Content

To move the network off of this zero equilibrium outcome, the site may implement a policy of generating a sufficiently large number of postings to attract more readers, which will eventually attract more writers. That is, the site may use site-generated content (SGC)

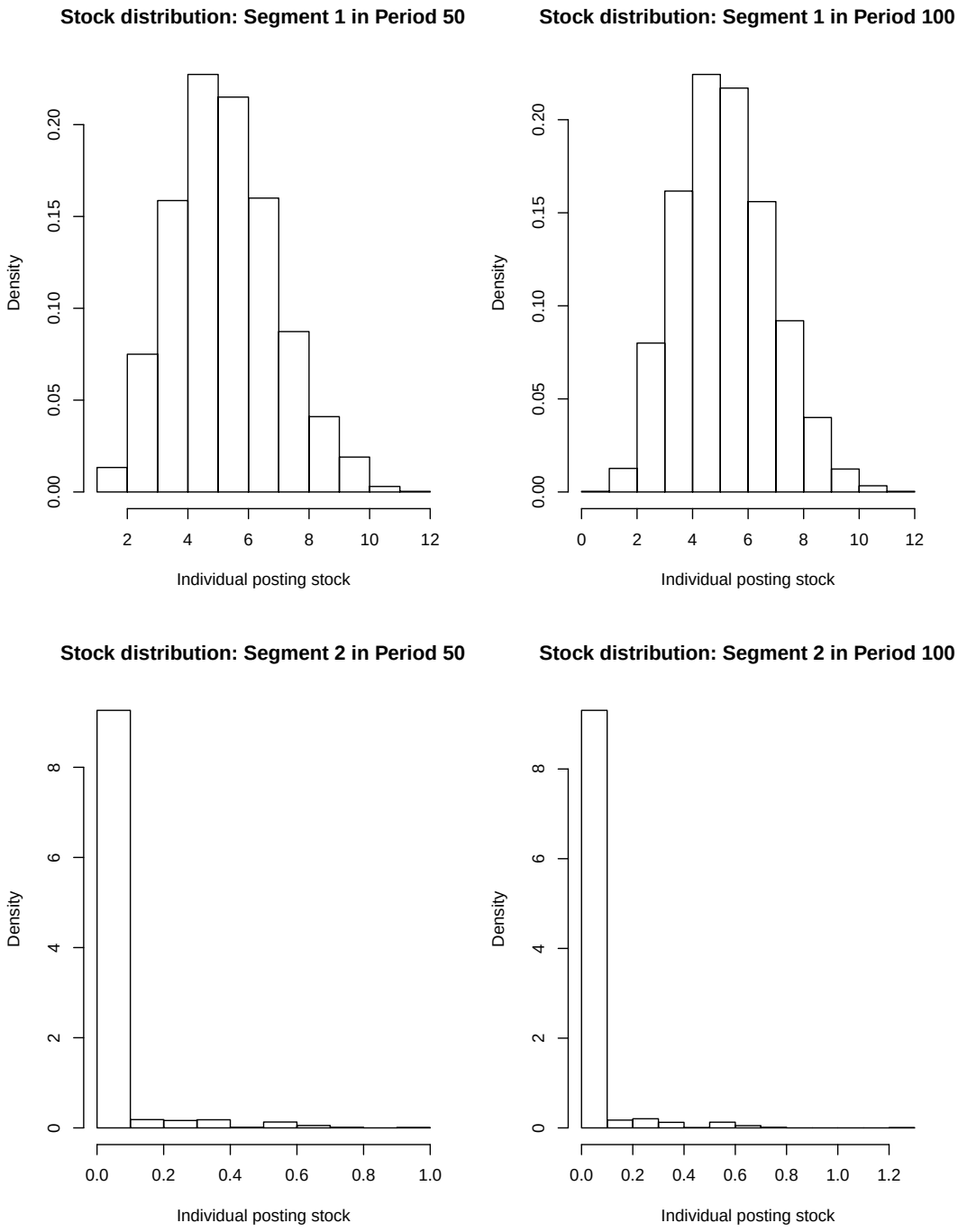

Figure 3: Distributions of individual user's posting stocks of the two segments defined in Section 4 in steady state.

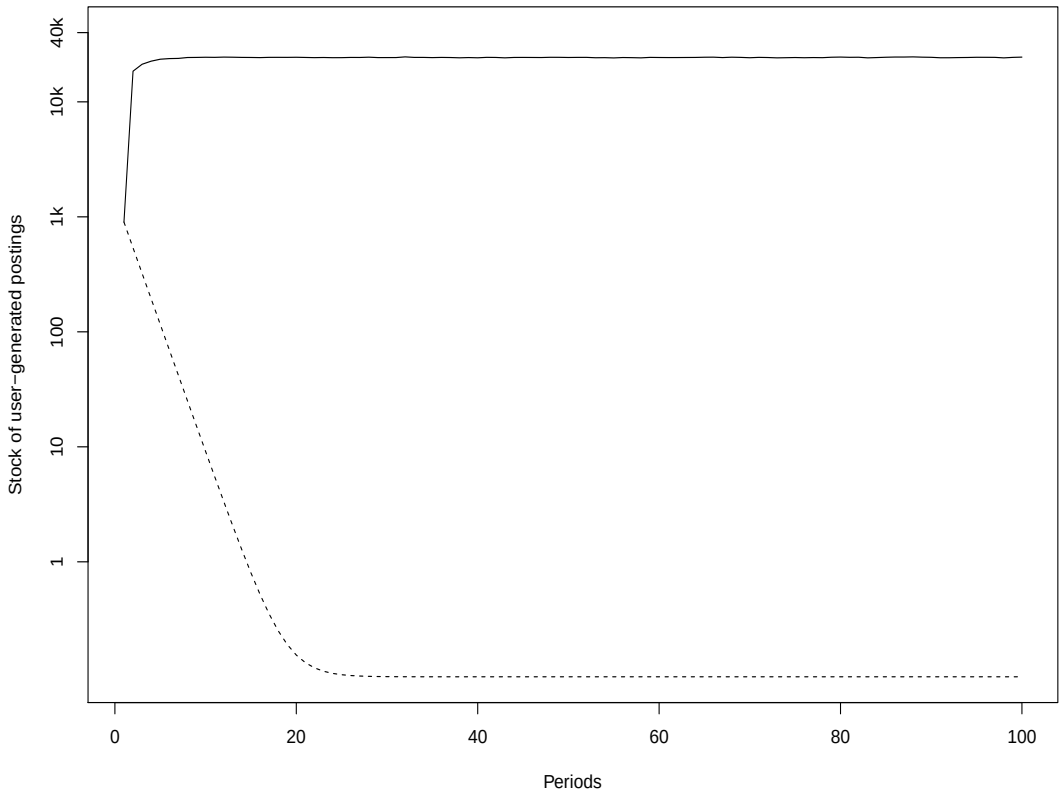

Figure 4: Convergence of the aggregate user-generated posting stock (UGC) to two different steady states from a common starting value when either i) the initial posting stock is random (solid curve) and 0.1 (dashed curve) as in Section 4.2.2or ii) the site-generated content (K_t^S) has the means equal to $20,000$ (solid curve) and $2,000$ (dashed curve) as in Section 4.2.3.

to "jump-start" and "bootstrap" user activity. As a simulation experiment, we choose a significantly higher level of SGC (with $\omega_0^S = 10000$, $\omega_1^S = 0.5$, $\sigma^S = 100$ and $K_{t=1}^S = 20000$) and test the model with the same low level of initial UGC (posting stock equal to 0.1 for all users in the two segments). We found the UGC converges to a different equilibrium (solid curve in Figure 4) which shows a much higher level of user activity.

4.2.4 Decay Parameter and Average Number of Postings per Person

The decay parameter ρ of forum postings implies two opposite effects on user posting activity. First, lower decay rate (higher ρ) means a post is more likely to be seen in the future, so a

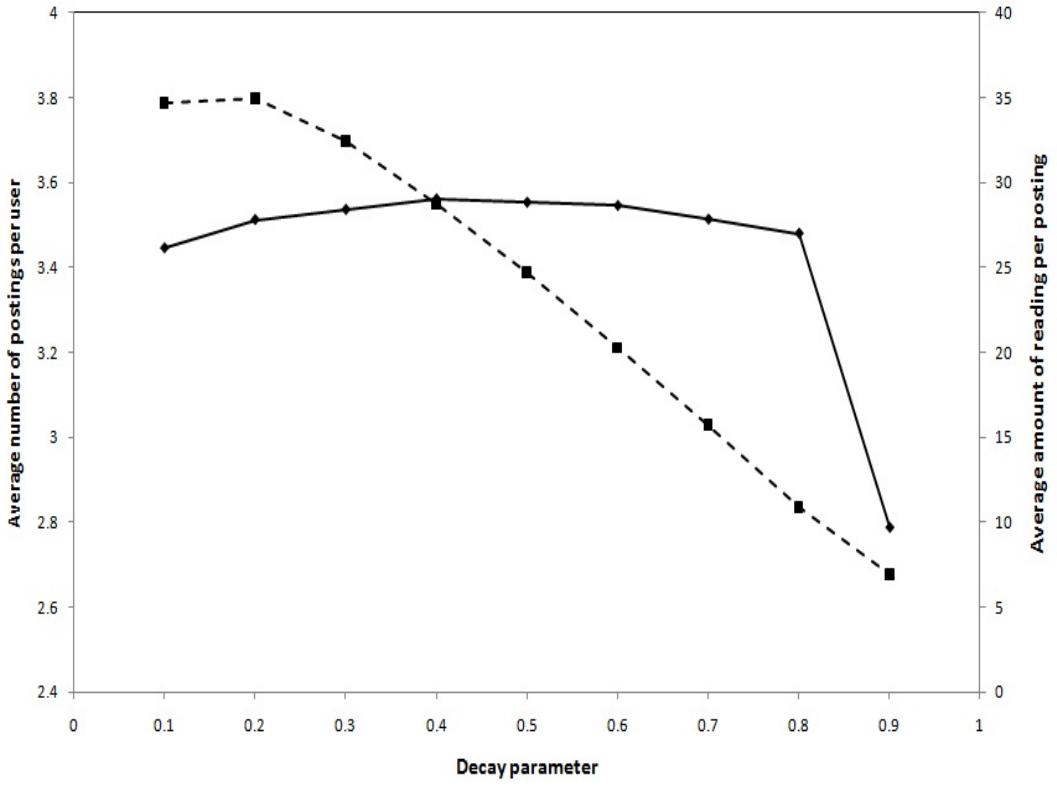

Figure 5: Average number of postings by individual users in steady state vs. the decay parameter ρ (solid curve) and the average reading per posting y_t vs. the decay parameter ρ (dashed curve).

user has the incentive to post more. This also raises content available for readers thereby increasing site participation. However, higher ρ makes posting more "durable" and hence increases the aggregate posting stock and decreases the rate of reading per posting (via competition for readers), which could cause a user to post less. The net effect of ρ is not clear directly from the utility function because a closed-form derivative of the utility with respect to the decay parameter cannot be easily derived. Therefore, we discretize the space of the decay parameter ($\rho \in [0, 1]$) to ten equally spaced grid points (0.1, 0.2, ..., 0.9) and simulate the content and reading given these values.

In Figure 5, we depict the relationship between the decay parameter and the average number of postings per period per user in Segment 1 (solid curve) based on the simulation

results. We also plot the relationship between the decay parameter and the average reading per posting y_t (dashed curve). From Figure 5, higher decay parameter will *ceteris paribus* cause lower average reading per posting thanks to the competitive effect of more durable stock postings. However, the average number of postings per user increases when the decay parameters ρ increases from 0.1 to 0.4 and decreases when ρ is above 0.5. This is due to the two opposite effects of ρ on user activity; attracting more readers to the site and increasing the overall reading of a given post over time.

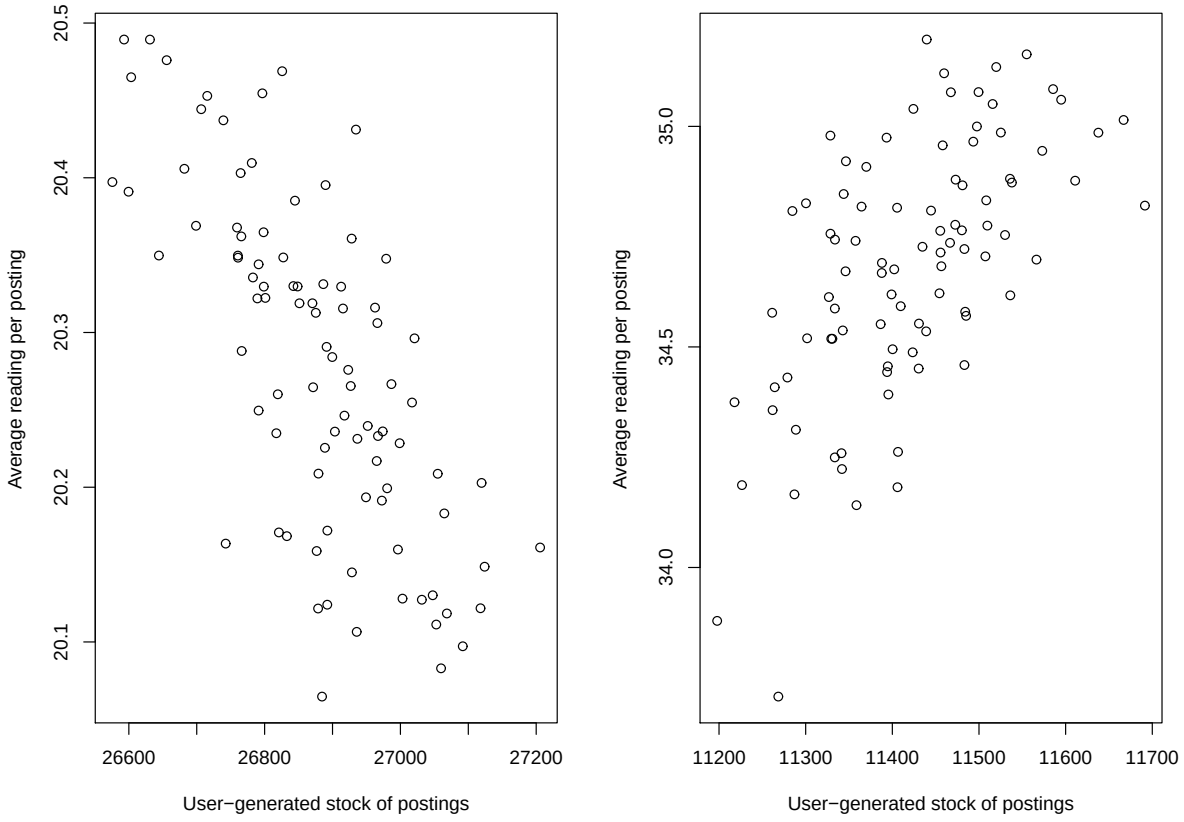

Figure 6: The relationship between average reading per posting and aggregate user-generated posting stock in equilibrium.

4.2.5 Indirect Network Effect of Aggregate UGC

The indirect network effect of the aggregate UGC on individual user's posting action is affected by the likelihood their post is read; that is, the numerator (aggregate reading) and denominator (aggregate postings) in equation (10). The numerator implies a greater likelihood of reading because more content, K_t^U, enhances the consumption experience. The denominator implies a competitive effect of K_t^U as content increasingly competes for users.

In Figure 6, we show two examples, one where y_t is decreasing in K_t^U and another where it is decreasing. The decay rate ρ is 0.6 for the first example and 0.1 for the second: all the remaining parameter values are identical in the two examples. We also find the relationship between y_t and K_t^Ucan switch sign if we adjust the ratio of population sizes of the two segments. Because the numerator is not a closed-form function of K_t^U, the conditions under which the network effect of K_t^U is positive are still unclear. We conjecture that positive indirect effect is more likely when there is a strong primary effect on site participation and that the negative indirect effect is more likely when the participation is already high.

5 Data

5.1 Data Overview

Our data come from a large Internet property devoted to a common interest like sporting events. The site includes a forum where persons can discuss various topics much like fans would discuss a sports team, its players or various games. We collect two months of forum participation data from October through November 2009, and use this as our basis of exploration for social engagement. The customer log files include the complete visit history for each registrant. The unit of observation is registrant-visit and indicate whether one reads or posts. We aggregate our data to a daily frequency and conduct our analysis at this periodicity, considering total reads and posts by each user on a daily basis inclusive of zeros.

Figure 7 plots the joint distribution of reading content generation and consumption, conditioned on non-zero reading (i.e., a site visit). The figure indicates that reading is more common than posting and days with higher posting rates tend to have higher reading rates. The large mode at zero is suggestive of the need to separately model the participation

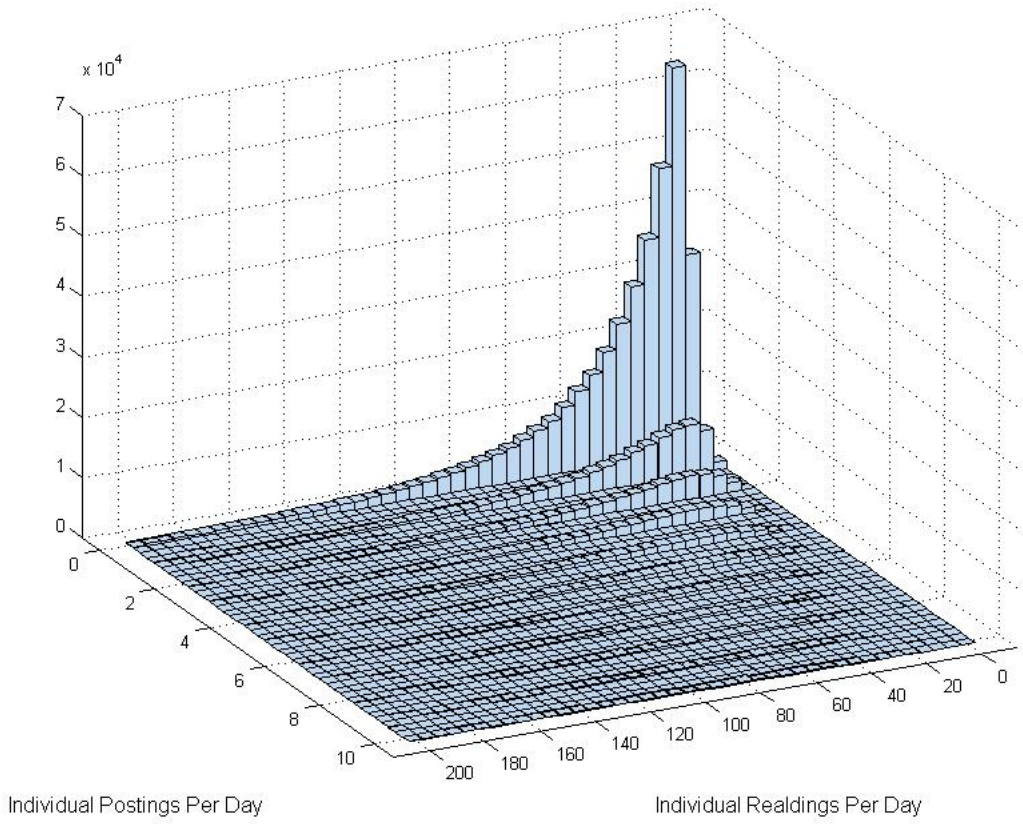

Figure 7: Joint Distribution of Reading and Posting

decision. The figure also indicates there is a fair amount of variation in the reading and posting behavior across observations.

5.2 Exploratory Analysis

To assess the potential for the presence of indirect effects and dynamics, we conduct a regression analysis. First, we consider reading. Recall, our model posited a positive link between aggregate content stock and individual reading. Hence, we regress the daily reads of individuals against content, using a Koyck formulation to capture posting stock effects (Clarke (1976)). Though we also include weekend (Thursday, Friday and Saturday) effects, we omit them from the following table to conserve space:

The contemporaneous effect of aggregate posts is significant as is the infinite horizon effect, which is given by 0.000033/(1-0.655). The exploratory regression suggests two things

Variable	Parameter Estimate	t-value	p-value
Aggregate Posting	0.000054	4.09	0.0001
Lag Reading	0.647	804.0	0.0001
n	898,139		
R^2	0.42		

Table 1: The Effect of Aggregate Posting Stock on Individual Reading

– first, posts have a stock effect consistent with Section 3.2.2 because the lag reading term is significant and second, that there is a positive indirect effect of posting stock on reading.

Variable	Parameter Estimate	t-value	p-value
Aggregate Reading Rates	165.36	29.07	0.0001
Aggregate Posting	0.0000106	3.67	0.0001
Lag Aggregate Posts	0.0000019	0.66	0.510
n	898,139		
R^2	0.0013		

Table 2: The Effect of Aggregate Reading Rates on Individual Posting

Next, we explore the indirect effect of reading rates and competitive effects of aggregate posting rates on the number of postings as discussed in Section 4.2.5. Table 2 reports the results of this regression. Consistent with our assumptions, the results suggest a strong effect of reading on the likelihood of posting. We also find a positive effect of aggregate posting on postings, suggesting that site participation effects dominate competitive effects (see 4.2.5). Overall, the exploratory analysis is consistent with the presence of indirect network effects and a stock effect for posts.

6 Estimation and Identification

6.1 Estimation

An efficient single-step estimation approach using maximum likelihood requires solving the dynamic optimization problem for every individual user and the rational expectations equilibrium for the aggregate reading and posting for each iteration of a nonlinear optimization program. The computational cost of this approach is considerable. To alleviate this problem, we design a two-step approach, which is a more computationally feasible estimation strategy. In this approach we first estimate the state transition equation for the aggregate UGC in

(27) and the reading-per-posting as a function of the UGC in (28) and then, in the second step, estimate the structural parameters in the individual reading, posting and site-visiting models.

In the second step, we impute these equations into the dynamic optimization model of posting. There are three estimation problems in the second stage: the reading, posting and site-visitation models. The reading model in equation (8) is estimated via maximum likelihood as detailed in Appendix E.1. The posting model estimation parallels Dubé et al. (2009), which is an MLE algorithm using mathematical programming with equilibrium constraints (MPEC). See Appendix E.2 for details. The site-visitation model in equation (25) is estimated as a binary choice model given the estimates of the parameters in the reading and posting models; this estimation is detailed in Appendix E.3.

6.2 Identification

Corresponding to the two-step estimation in Section 6.1, there are two sets of parameters to identify in the model: (i) the parameters in the aggregate models (27) and (28) and (ii) the structural parameters in the individual utility and cost functions of reading, posting and site-visitation.

We first consider the system of equations in the first stage estimation model of aggregate state levels. Site-generated content K_t^S is set to be zero because the website does not generate a substantial number of postings (only dozens of postings a day at most and zero in some days) in our data. Because the UGC and its corresponding stock K_t^U is generated endogenously with the rate of reading per posting y_t, the identification of these aggregate state transition models follows from the exclusion restrictions due to the lag posting stock K_{t-1} in equation (27), which is similar to using lagged prices in Nair (2007). First, we note that the lag stock is correlated with current stock. This claim follows from the definition of the individual stock variables which are comprised of decayed individual posting stock k_{it} and any current period additions to content a_{it}; this decay process leads to a strong auto-correlation for the aggregate K_t that increases with the stock decay parameter, ρ. Indeed, the auto-correlation parameter ω_1^U in equation (27), which is a reduced-form AR(1) model for the aggregate K_t, is estimated to be 0.93 by the data. Next, we argue that the lag stock is

not correlated with current period reading rates y_{it}. Based on the individual reading model in Section 3.2, the average amount of reading per posting y_t is the aggregate r_{it}'s divided by K_t and the r_{it} is aslo a function of the current stock K_t. Hence, y_{it} should be conditionally independent of K_{t-1}^U given K_t^U, which is reflected by equation (28).

In the second stage, the parameters in the individual reading model are identified by the data. In equation (9), the mean level of reading is equal to $(\alpha_1 - \kappa_{1it} - \zeta_i)/(\alpha_2/K_t + \kappa_{2i})$, which implies identical value if we multiply both the numerator and the denominator by a common factor. For identification, we therefore fix α_2 to be one. The effects $\bar{\zeta}_j$ for the heterogeneous cost in the J segments are identified by the cross section of mean reading rates across users and the fixed effects for the week-days are identified by the variation of reading in individual time series. Given the estimated state transition for the aggregate variable K_t, the individual-level posting model is a single-agent dynamic discrete choice model which is nonlinear in the utility and cost parameters. These parameters are identified by the panel structure of the posting data. Note that the aggregate states are considered to be exogenous in the individual level reading and content generation models because, as the size of the reading and posting populations become large, the expectation of the sum of the individual level shocks tends to zero and becomes independent of the individual-specific shocks. Lastly, the site-visitation model is a simple nested logit model where the scale parameters μ_0 and μ_1 are identified by the panel structure of the site-visitation data.

7 Results

7.1 Initialization of Posting Stock

As indicated in Section 3.2.2, the posting stock is incumbent upon the decay rate of a post. We estimate the exogenous posting decay parameter ρ using auxiliary data collected by the Internet site regarding when a sample set of the site's posts were visited by its users. The decay in the number of users clicking on these posts over time is informative about their durability. From these data, we consider a random sample of 474 postings posted on the forums in the first week of sampling period.

The decay parameter is identified by the ratio of the times that a posting is read in

periods t and $t+1$. Note that this ratio is independent of the endogenous average amount of reading per posting y_t. Under the exponential decay assumption, this ratio equals the decay rate in the amount of reading per posting (the ratio of reading per posting in period t divided by reading per posting in $t+1$). By stacking all observations of this ratio across users and periods, generalized least squares and can therefore be used to infer the post decay. We use feasible GLS to control for high variation in this ratio for observations in excess of 10 days after a post (because there are few reads after 10 days, this ratio becomes less reliable). In addition, we control for potential seasonal effect due to the day of the week. The resulting estimate for the mean decay is 0.737, which implies that 90% of the post's stock is depleted after one week.

Note that we do not observe individuals' initial posting stocks in the first week of the data as there is no history of posts prior to the initial week. Hence, using this posting stock decay estimate, the individual posting stock is computed by setting the initial stock at zero and recursively applying equation (2) using the 61-day posting data repeatedly until the individual's posting stock reaches a steady state. The individual's steady state is then re-used as the initial posting stock to calculate the individual posting stock for the 61-day data. We adopt this practice because the users in our sample have been using the forum for long time prior to the sampling period, hence their posting stocks are likely to have reach the steady state (with daily random variation) at the inception of our data. We similarly compute the aggregate stock K_t^U for the same sample stock. The site-generated content K_t^S is set to zero because there are too few K_t^S relative to the K_t^U (about 0.02%).

7.2 Approximate Aggregation Results

Section 3.7 outlines the aggregate state transition model that captures the rational expectations process. The estimation results for the AR(1) model in (27) and (28) are reported in Table (3). The results provide evidence of strong auto-correlation ($\omega_1^U = 0.93$) for the aggregate stock. The rate of reading per posting is an increasing function of aggregate stock ($\omega_1^y = 5.43 \times 10^{-6}$ is statistically significant), which implies the positive indirect network effect of posting on site participation exceeds the negative competitive effect for our forum data. The week-day effects for Monday and Tuesday in model (27) are not significantly dif-

Model	AR(1) for UGC stock	Reading-per-posting
Intercept ω_0^U or ω_0^y	$[1945, 55909]$	$6.02\ [4.14, 7.90]$
Lag UGC stock ω_1^U	$0.93\ [0.86, 0.99]$	–
Current UGC stock ω_1^y	–	$5.43 \times 10^{-6}\ [0.36, 10.5] \times 10^{-6}$
Weekend effect ω_{0t}^U or ω_{0t}^y	$-6922\ [-8802, -5043]$	$-0.55\ [-0.69, -0.42]$
Residual R^2	0.89	0.51

Table 3: Estimation Results for Aggregate Posting Stock Transition Equation and Rate of Reading-per-Posting Equation (with 95% confidence intervals in brackets)

ferent from Sunday, whereas the effects for Wednesday, Thursday, Friday and Saturday are significantly negative, which implies lower posting activity for these days of a week. All the week-day effects in model (28) are significantly negative, which means lower reading activity for these days.

7.3 Individual-level Model Results (Posting, Reading and Site Visitation)

We randomly select a sample of 600 users users to estimate the individual-level model. The amount of reading and number of postings for each individual in the sample are recorded for 61 days from October 1st to November 30th, 2009. If both reading and posting are zero for a user in a certain day, we conclude the user does not visit the site that day.

7.3.1 Estimates of Utility and Cost Parameters in Posting and Reading

Table 4 reports parameter estimates for the posting and reading models assuming two segments of users. for a discount parameter of $\beta = 0.98$.[14] The two segments are specified to share a common posting utility parameter, γ, in equation (12) but differ with respect to their posting costs, $\bar{\xi}_j$, in equation (13) as heterogeneity in costs and utilities are not separately identified. Likewise, the two segments in the reading model share a common utility parameter, α_1, but differ with respect to linear marginal cost parameter, $\bar{\zeta}_j$, because heterogeneity in costs and utilities are also not separately identified.

Comparing the two groups, the second segment is slightly smaller in size and evidences higher reading and posting costs; hence, this group of users read less often and rarely posts

[14]We also test three segment of users. However, the BIC for the three-segment model is higher than the two-segment model.

Parameters	First Segment Frequent Users	Second Segment Light Users
Posting Model		
Utility coefficient γ	0.85*[0.79, 0.88]	
Cost coefficient $\bar{\xi}_j$	1.17* [0.74, 1.32]	7.64* [3.00, 8.90]
Weekend effect $\bar{\tau}_{jt}$	0.026* [0.01, 0.57]	0.73* [0.09, 1.30]
Reading Model		
Utility coefficient α_1	11.51* [3.66, 17.56]	
Linear cost coefficient $\bar{\zeta}_j$	−7.88* [−1.59, −12.64]	7.88* [1.59, 12.64]
Weekend cost effect κ_{1t}	−0.16[−1.23, 0.88]	0.041[−0.24, 0.32]
Quadratic cost coefficient $\bar{\kappa}_{2j}$	0.36 [0.07, 0.76]	0.16 [0.04, 0.21]
Site Visitation Model		
Intercept, μ_{0j}	1.15 [0.24, 4.91]	0.70[0.23, 1.11]
Reading scale parameter μ_1	0.10 [0.032, 0.16]	
Gumbel scale parameter η	0.46 [0.12, 0.80]	
Heterogeneity		
Segment size	59.7% [41.4%, 65.5%]	40.3% [34.5%, 58.6%]

Table 4: Estimation Results for Utility and Cost Parameters in Posting and Reading

content. Hence, we denote them "light users." Also of note, the weekend effect $\bar{\tau}_{jt}$ in the posting cost function is positive, so the users tend to post less on a weekend.

7.3.2 Site-Visitation Estimates

As indicated in Table 4, the intercept for the frequent users is higher than light users; all else equal the frequent users visit more. However, the difference is not statistically significant, suggesting the observed difference in visitation rates across segments is primarily due to expected reading and writing utility upon visitation, and that the time invariant unobserved factors add little to distinguish the visitation behavior across segments. We estimate a common set of scale parameters because it is not clear why these scale factors should differ across segments (of note, we find no significant difference when we estimate them separately).

7.4 Policy Simulations and Comparative Statics

In this section, we consider the policy ramifications of our model. One concern involves how the site should manage its own content development strategy to enhance traffic. Second, we consider the choice of whether to invest in new contributors or new readers. Our model, by measuring network effects, affords insights into which of the two strategies generates a

higher marginal effect on overall site usage. A third policy experiments concerns increasing the durability of posts and, in a fourth, we consider the potential for self-fulfilling prophecies to assess whether erroneous beliefs can yield different participation outcomes in the steady state.

7.4.1 Site Generated Content Strategies

As indicated in Figure 6, the ex ante effect of additional content on posting is ambiguous. On one hand, there is a competitive effect that lowers reading likelihood of the other posts. On the other hand, increased content can generate more readership, thereby increasing the utility of posting and the resultant posts. We consider this trade off explicitly and intend to make recommendations regarding the site's participation levels. It is worth noting that these levels are currently negligible[15] and that the site management is particularly interested in the outcome of this analysis.

Given that the current net network effect of UGC estimated from the data is positive, we postulate that smaller amount of SGC will attract more readers to the forum, which will lead to higher utility for posting and consequently greater amount of UGC. However, higher SGC will make the competitive effect dominant and cause the endogenous net network effect negative eventually, which will reduce UGC.

We simulate user-generated content (UGC) in the rational expectations equilibrium by manipulating the site-generated content in the current equilibrium; that is we consider incremental site content over the long run average level of user content. Because the site usually hires users of the forum to write postings, we further assume forum readers view and respond to the site-generated content in the same way as to the user-generated content. The resulting percentage change in average user-generated postings and number of visitors in equilibrium over a 70-day simulated sequence versus the levels of SGC is plotted in Figure 8.

Figure 8 demonstrates that SGC initially increases both the number of visitors and the user-generated content. As the site content increases further to about 7%, it begins to reduce the amount of UGC as the competitive effect of postings begin to predominate.[16] At that

[15] In the current data, site posts are less then .2% of total daily posts.

[16] The seemingly oscillating behavior of the curves is due to sampling errors, as the mean number of postings is only the average of a 70-day simulated sequence.

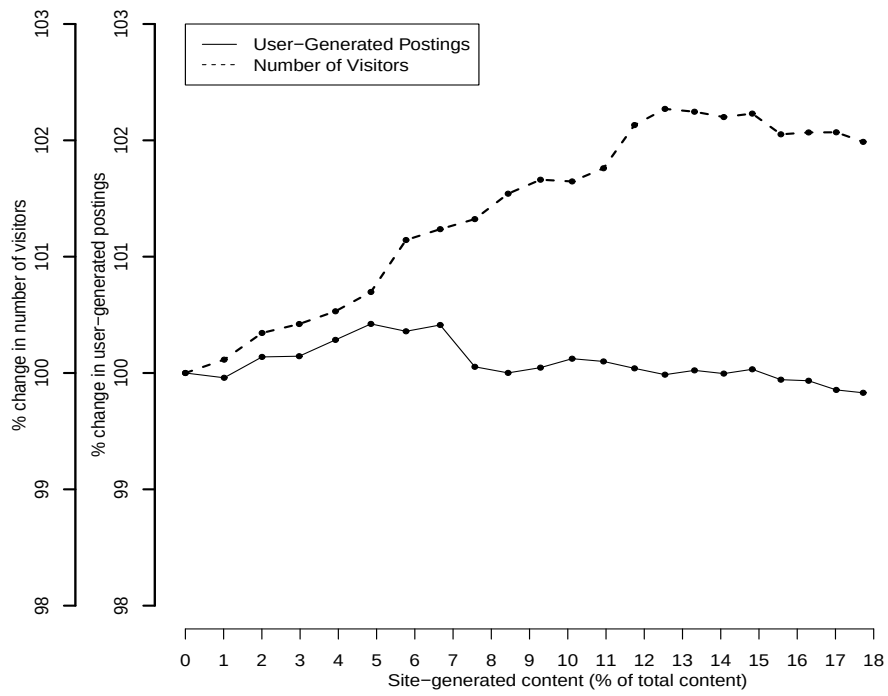

Figure 8: Effect of site generated content strategies on user-generated content and number of site visitor

point, the increment in UGC due to the site-generated content is only about 0.4%. As the site content increases further to 12%, it begins to reduce the number of visitors in response to the continued decrease in the number of user posts . At this point, the incremental number of visitors at this point reaches its maximum of 2.2%. Whether the strategy is profitable depends on the relative costs of generating the content and the advertising revenue generated by having 2.2% more users.

7.4.2 Lowering User Generation and Reading Costs

Pursuant to the consideration of whether to invest in increasing reading and posting, we consider the marginal effect of reading and posting costs. These costs can be lowered by changes in site design, emails to users or possibly incentives. Specifically, we intend to compute the effect of a 10% decrease in reading and posting costs on overall site traffic

(number of visitors per day) and the number of postings. Such a simulation will generate insights into where the greatest efficiency in investment might be. We find reducing the cost of *posting* for both segments of forum users will increase the aggregate number of postings by 18.9% and the number of visitors by 3.8%. Reducing the cost of *reading* by 10% will increase the aggregate number of postings by 7.9% and the number of visitors by 11.3%. Hence, we conclude that strategies reducing reading cost, such as making the forum more easily accessible to readers such as providing forum access applications for smart phone users, are the most efficient ways to promote site traffic as long as the expense of reducing reading costs is sufficiently similar to that of reducing posting costs.

If we ignore the rational expectations equilibrium for these policy simulations, the aggregate state transition for UGC will be spurious when the costs of posting or reading are reduced by 10%. When the cost of *posting* is reduced by 10% and the rational expectations equilibrium is ignored, the simulation predicts 21.1% gain in the aggregate number of postings and 4.1% increase in the number of site visitors. This is because ignoring the rational expectations equilibrium underestimates both the competing effect of postings and the rate of reading per posting. Hence, the model over-estimates UGC by 11% and the number of visitors by 7%. When the cost of *reading* is reduced by 10% and the rational expectations equilibrium is ignored, the model predicts 4.4% gain in the the number of postings and 8.4% increase in the number of site visitors. Ignoring the rational expectations equilibrium underestimates the increase in postings by 44% when there are more visitors, which in turn underestimates the increase in the number of visitors by 26%.

7.4.3 Extending Post Durability

In Footnote Section 8, we demonstrated that increasing the durability of postings gives users incentive to post more. However, increasing durability can also decrease postings owing to the increasing competitive effect of past postings. To test whether the forum can promote posting activity and increase the number of visitors by extending the durability of postings, we raise the decay parameter ρ from the estimated 0.737 to 0.85, which approximately doubles the expected life-time of any given postings (the 90% decay interval increase from one to two weeks). Results indicate that raising ρ to 0.85 will increase user generated postings

by 22.8% and the number of visitors by 9.1%. Because extending the durability of postings can be achieved by employing a better search engine and/or improving the website layout, which often incurs only a one-time sunk cost, it may be a very fruitful strategy to enhance participation at the site.

7.4.4 Self-fulfilling Prophecies

Owing to the formation of beliefs regarding aggregate state transitions as indicated in equations 26 - 28, content generation and reading decisions are incumbent upon future beliefs. Of interest is the possibility that these beliefs become self-reinforcing. This issue can be explored by shocking these beliefs in the short-term (by varying the initial states and the variances in the state transition equations) and the long-term (by varying the regression coefficients in the state transition equations) seeing how the evolution of content generation and consumption change relative to a situation where the beliefs are initially inconsistent with the long-term behaviors.

In order to test whether shocking short-term beliefs can lead to different long-term behaviors, i.e., converging to different equilibria of the model, we reset the initial belief about the aggregate user-generated posting stock to 5%, 25%, 50%, 150%, and 200% of the observed actual stock and simulate the rational expectation equilibrium following the algorithm in Section 3.7. We find all these simulations converge to the same equilibrium which has the same levels of mean UGC and number of visitors as in the observed data. We also reset the initial belief about the variance in the state transition equation for the aggregate UGC to 25%, 50%, 150%, 200% and 300% of the value estimated from the real data. All these simulations again converge to the original equilibrium. Hence, we conclude that shocking short term beliefs will not lead to self-fulfilling behavior.

To evaluate whether erroneous long-term beliefs about the transition rule of aggregate UGC can lead to different equilibrium, we set the initial value for the auto-regressive coefficient ω_1^U in equation (27) to 0.1, 0.2,...,0.9 and simulate their corresponding equilibria. We find they always converge to the same equilibrium in which the auto-regressive coefficient ω_1^U is 0.84. Therefore, erroneous long-term beliefs about the transition rule will not lead to self-fulfilling behavior. Note that the rational expectations equilibrium in our model is

similar to that by Krusell and Smith (1998), who also found the absence of self-fulfilling behavior in their model.

8 Conclusions

Recent advances in technology and media have enabled user generated content sites to become an increasingly prevalent source of information for consumers as well as an increasingly relevant channel for advertisers to reach users of these sites. Hence, the factors driving the use of these networks is of a topical concern to marketers. In this paper, therefore, we consider how content, readership and site policy drive the evolution of content and readership on these sites.

Given our goal is to develop prescriptive and theoretical insights regarding user engagement on user generated content platforms, we build upon the existing literature on social participation by developing a dynamic structural model to explore these effects. Individual reading behavior is developed from a model of information search that relates reading to the overall level of content on the site. Individual content generation is assumed to reflect the utility that participants receive from the number of others reading the posts. Underpinning these two behaviors are users' beliefs regarding how the aggregate amount of content and readership on the platform evolve. These beliefs stem from the rational expectations equilibrium model whereby the evolution of aggregate reading and content states are assumed to be consistent with the aggregation of individual level reading and contribution decisions across the population.

Our paper makes several contributions. On a methodological front, we develop a dynamic structural model of user generated content. To our knowledge, it is also the first paper in marketing to apply the approximate aggregation approach of Krusell and Smith (1998), which facilitates the computation of rational expectations equilibrium in the face of a large number of heterogeneous agents. This approach could prove useful in other contexts wherein firms face heterogeneous consumers. For example, heterogeneous learning about new consumer products can affect how prices evolve, and consumer may anticipate and react to such changes Narayanan and Manchanda (2009). Initial estimates of our model of UGC demonstrate that the indirect network effect or aggregate reading on posting and aggregate posting on reading

are both significant.

On a theoretical dimension, we explore the tipping effects and self fulfilling prophecies in the context of two sided network wherein one side involved content creation and one involved content consumption. We find that the potential exists for multiple equilibria depending upon whether initial usage can cross a sufficient threshold to attract participation; users will not visit the site to read if there is no material and users will not post if there are no readers. Of future interest, this approach can be applied to assess the formation or dissolution of similar networks, such as academic journals (readers and authors), social media sites, blogs and so forth. Another theoretical insight is that user and site content can serve as strategic complements or substitutes depending on whether the primary demand effect of content (attracting more users) dominates the secondary demand effect (splitting readers). An analogous argument can be constructed for past and current posts as their durability increases.

On a substantive domain, we consider a number of policy prescriptions to advise the sponsoring site. First, we consider the role of their own content on user participation. On the one hand, site posts attract more readers, thereby growing the network. On the other hand, these posts are competitive with other users' posts for reader attention. Overall, we conclude that the former effect predominates and the site can increase visitation by 2.2% by increasing content by 12%. Beyond this point, the sites posts crowd user posts leading to a decrease in posts and visitors. In addition, we explore the relative effect of reducing user cost of participation and contrast the relative effect of reading and content generation costs, finding that lower posting costs has a more substantial effect than lowering reading costs. To the extent that cost mitigation strategies are equivalently expensive to implement, the posting cost reduction should be the first considered. Further, we find that the durability of posts has an effect analogous to site generated contents inasmuch as old posts can serve as strategic substitutes or complements (they also lower the effective costs per post). Overall, we find that increasing the durability of posts is one of the more effective strategies for enhancing site participation and engagement.

Several opportunities for extensions present. First, the site we consider is the largest forum by market share on the topic it covers much like Youtube is for videos. In practice, the

potential for competition exists in other context and extending our work to a duopoly context would be of interest Zhang and Sarvary (2011). Second, content is not homogeneous and it would useful to extend our model to capture heterogeneity in content in order to explore which information is most relevant in increasing site engagement. Related, the potential exists that certain lead content creators generate large followings (such as on Twitter) and measuring the effect of lead users is of practical interest. While sites generally consider such participation to be positive, it is also possible for the content to compete with others and actually reduce site participation. In sum, we hope that our research will lead to additional innovations in both user generated content and the application of the rational expectations equilibrium theory in marketing.

References

Albuquerque, Paulo, Polykarpos Pavlidis, Udi Chatow, Kay-Yut Chen, Zainab Jamal, Kok-Wei Koh, Andrew Fitzhugh. 2010. Evaluating Promotional Activities in an Online Two-Sided Market of User-Generated Content. *SSRN eLibrary* .

Ansari, Asim, Oded Koenigsberg, Florian Stahl. 2011. Modeling multiple relationships in social networks. *Journal of Marketing Research* **48**(4) 713 – 728.

Bughin, Jacques R. 2007. How companies can make the most of user geneated content. *McKinsey Quarterly* 1–4.

Bulte, Christophe Van Den. 2007. *Social Networks and Marketing*. Marketing Science Institute, Cambridge MA.

Chevalier, Judith A., Cina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research* **43**(3) 345–354.

Choi, Ki-Hong, Choon-Geol Moon. 1997. Generalized extreme value model and additively separable generator function. *Journal of Econometrics* **76**(1-2) 129 – 140.

Clarke, Darral G. 1976. Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research* **13**(4) pp. 345–357.

Dellarocas, Chrysanthos. 2006. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science* **52**(10) 1577–1593.

Duan, W., B. Gu, A. B. Whinston. 2008. Do online reviews matter?: An empirical investigation of panel data. *Decision Support Systems* **45**(4) 1007–16.

Dubé, Jean-Pierre, Günter J. Hitsch, Puneet Manchanda. 2005. An Empirical Model of Advertising Dynamics. *Quantitative Marketing and Economics* **3** 107–144, 10.1007/s11129-005-0334-2.

Dubé, Jean-Pierre H., Günter J. Hitsch, Pradeep K. Chintagunta. 2010. Tipping and concentration in markets with indirect network effects. *Marketing Science* **29**(2) 216–249.

Dubé, Jean-Pierre H., Jeremy T. Fox, Che-Lin Su. 2009. Improving the Numerical Performance of Blp Static and Dynamic Discrete Choice Random Coefficients Demand Estimation. *SSRN eLibrary* .

Geweke, John, Micheal P. Keane. 1997. Mixture of normals probit models. *Research Department Staff Report 237, Federal Reserve Bank of Minneapolis* .

Ghose, Anindya, Sang Pil Han. 2010. An Empirical Analysis of User Content Generation and Usage Behavior in the Mobile Internet. *SSRN eLibrary* .

Ghose, Anindya, Sang Pil Han. 2011. A Dynamic Structural Model of User Learning on the Mobile Internet. *SSRN eLibrary* .

Hartmann, Wesley R. 2010. Demand estimation with social interactions and the implications for targeted marketing. *Marketing Science* **29**(4) 585–601.

Hennig-Thurau, Thorsten, Kevin P. Gwinner, Gianfranco Walsh, Dwayne D. Gremler. 2004. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing* **18**(1) 38–52.

Hofstetter, Reto, Scott K. Shrivery, Harikesh S. Nair. 2010. Social ties and user generated content: Evidence from an online social network. *Working Paper, Stanford University* 1–.

Huang, Yan, Param V. Singh, Anindya Ghose. 2011. A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media. *SSRN eLibrary* .

Iyengar, Raghuram, Christophe Van den Bulte, Thomas W. Valente. 2010. Opinion leadership and social contagion in new product diffusion. *Marketing Science* mksc.1100.0566.

Kamakura, Wagner A., Gary J. Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research* **26**(4) pp. 379–390.

Katona, Zsolt, Peter Pal Zubcsek, Miklos Sarvary. 2011. Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research* **48**(3) 425 – 443.

Krusell, Per, Anthony A. Smith. 1998. Income and wealth heterogeneity in the macroeconomy. *The Journal of Political Economy* **106**(5) pp. 867–896.

Lee, Donghoon, Kenneth I. Wolpin. 2006. Intersectoral labor mobility and the growth of the service sector. *Econometrica* **74**(1) 1–46.

Miranda, Mario S., Walter D. Fackler. 2002. *Applied Computational Economics and Finance*. The MIT Press, Cambridge, MA.

Nair, Harikesh. 2007. Intertemporal price discrimination with forward-looking consumers: Application to the us market for console video-games. *Quantitative Marketing & Economics* **5**(3) 239 – 292.

Nair, Harikesh S, Puneet Manchanda, Tulikaa Bhatia. 2010. Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. *Journal of Marketing Research* **47**(5) 883 – 895.

Narayanan, Sridhar, Puneet Manchanda. 2009. Heterogeneous learning and the targeting of marketing communication for new products. *Marketing Science* **28** 424–441.

Nardi, Bonnie A., Diane J. Schiano, Michelle Gumbrecht, Luke Swartz. 2004. Why we blog. *Commun. ACM* **47** 41–46.

Nov, Oded. 2007. What motivates wikipedians? *Communications ACM* **50** 60–64.

Rust, John. 1987. Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society* (5) 999–1033.

Rust, John. 1994. *Structural Estimation of Markov Decision Processes*. Amsterdam: Elsevier Science.

Stephen, Andrew T., Oliviet Toubia. 2010. Deriving value from social commerce networks. *Journal of Marketing Research* **47**(2) 215–228.

Stigler, George J. 1961. The economics of information. *The Journal of Political Economy* **69**(3) pp. 213–225.

Su, Che Lin, Kenneth L. Judd. 2010. Structural estimation of discrete-choice games of incomplete information with multiple equilibria. *Proceedings of the Behavioral and Quantitative Game Theory: Conference on Future Directions*. BQGT '10, ACM, New York, NY, USA, 39:1–39:1.

Yao, Song, Carl F. Mela. 2008. Online Auction Demand. *Marketing Science* **27**(5) 861–885.

Zhang, Kaifu, Theodoros Evgeniou, V. Padmanabhan, Emile Richard. 2011. Content contributor management and network effects in a ugc environment. *Marketing Science* **Forthcoming** 1–.

Zhang, Kaifu, Miklos Sarvary. 2011. Social media competition: Differentiation with user generated content. *Working Paper, INSEAD* 1–.

Appendix

A The Utility of Reading

Assume the quality of a given content item has a uniform distribution on a closed interval $[L, U]$, where $U > L \geq 0$ and $U - L$ is the length of the support of the uniform distribution. Note the mean is $(U + L)/2$ and the variance is $(U - L)^2/12$. The qualities of the $\tilde{K}_t$ postings noticed by individual i, denoted as $Q_1, \ldots, Q_{\tilde{K}_t}$, are iid from $Unif[L, U]$.

We assume the individual reads the postings according to their quality ranking, then the marginal utility gains measured by incremental quality from an additional posting is an increasing function of the total stock postings $\tilde{K}_t$. Let the qualities of $\tilde{K}_t$ postings be ranked as their order statistics $Q_{[1]} \leq Q_{[2]} \leq \cdots \leq Q_{[\tilde{K}_t]}$. $Q_{[k]}$ has the following distribution:

$$Q_{[k]} \sim \frac{\tilde{K}_t!}{(k-1)!\left(\tilde{K}_t - k\right)!} \left(\frac{q-L}{U-L}\right)^{k-1} \left(\frac{U-q}{U-L}\right)^{\tilde{K}_t - k} \frac{1}{U-L} \tag{A1}$$

Note this is linear transformation from a Beta distribution. That is $\left(Q_{[k]} - L\right)/(U - L)$ has a $Beta(k, \tilde{K}_t + 1 - k)$ distribution. Therefore we have

$$E\left(Q_{[k]}|\tilde{K}_t\right) = (U - L)\frac{k}{\tilde{K}_t + 1} + L. \tag{A2}$$

If individual i select to reads r_i highest quality postings, the expected utility given $\tilde{K}_t$ is

$$\begin{aligned} u(r_i) &= E\left(\sum_{k=\tilde{K}_t - r_i + 1}^{\tilde{K}_t} Q_{[k]}|\tilde{K}_t\right) = \sum_{k=\tilde{K}_t - r_i + 1}^{\tilde{K}_t} \left\{\frac{(U-L)\,k}{\tilde{K}_t + 1} + L\right\} \\ &= (U - L)\left\{\frac{\tilde{K}_t + 1/2}{\tilde{K}_t + 1} r_i - \frac{1}{\tilde{K}_t + 1}\frac{r_i^2}{2}\right\} + Lr_i. \end{aligned} \tag{A3}$$

We can approximate the realized stock $\tilde{K}_t$ with its expected value K_t thanks to the law of large numbers and $\tilde{K}_t$ being very large (over 100,000 in our data) . We have

$$u(r_i) = (U - L)\left\{\frac{K_t + 1/2}{K_t + 1} r_i - \frac{1}{K_t + 1}\frac{r_i^2}{2}\right\} + Lr_i. \tag{A4}$$

and the marginal utility of reading

$$\frac{d}{dr_i} u(r_i) = (U - L)\frac{K_t + 1/2}{K_t + 1} + L - \frac{U - L}{K_t + 1} r_i, \tag{A5}$$

is an increasing function of K_t. We can reparametrize $\alpha_1 = U$ and $\alpha_2 = U - L$ and define

$$u(r_i) = \frac{\alpha_1 K_t + (\alpha_1 - \alpha_2) + \frac{1}{2}\alpha_2}{K_t + 1} r_i - \frac{\alpha_2}{K_t + 1}\frac{r_i^2}{2}. \tag{A6}$$

Note the utility of reading can be further simplified when K_t is a large number using the approximation $K_t + 1 \approx K_t$ and $\left[K_t + (\alpha_1 - \alpha_2)/\alpha_1 + \alpha_2/2\alpha_1\right]/(K_t + 1) \approx 1$. In that case, we have

$$u(r_i) \approx \alpha_1 r_{it} - \frac{\alpha_2 r_{it}^2}{2K_t}. \tag{A7}$$

B Aggregate Reading

Here we show the expected amount of reading per posting y_t define in equation (10) can be closely approximated by the observed amount of reading per posting. By summing equation 9, the expected amount of reading of a given user, we obtain the aggregate expected amount of reading by all users,

$$R_t = E\left(\sum_{i=1}^{M} n_{it} r_{it}^*\right) = \sum_{i=1}^{M} E\left(n_{it} r_{it}^*\right), \tag{A8}$$

where M is the total number of users.[17] When we apply the latent segment model, the expected amount of reading of any user i is

$$\begin{aligned} E\left(n_{it} r_{it}^*\right) &= E\left[E\left(n_{it} r_{it}^* | n_{it}, \zeta_i\right)\right] \\ &= E\left[n_{it} E\left(r_{it}^* | n_{it}, \zeta_i\right)\right] \\ &= \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) E\left(r_{it}^* | \bar{\zeta}_j, n_{it} = 1\right) dF\left(s_{it}\right), \end{aligned} \tag{A9}$$

where $F(s_{it})$ is the stationary distribution of the state variables s_{it} and $p(n_{it} = 1|s_{it})$ is the probability that the user i visits the site at period t defined in Section 3.5. By substituting A9 into A8, we have

$$R_t = \sum_{i=1}^{M} \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) E_\nu\left(r_{it}^* | \bar{\zeta}_j, n_{it} = 1\right) dF\left(s_{it}\right) \tag{A10}$$

$$= M \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) \frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta}_j}{\alpha_2 / K_t + \kappa_{2j}} dF\left(s_{it}\right), \tag{A11}$$

[17] As the number of registered users does not change vary by more than 3% over the duration of our data, we treat the market size, M, as fixed over time in our model. That said, overall traffic can increase when the likelihood of visiting a site increases.

and obviously $R_t/M = E\left(n_{it} r_{it}^*\right)$.

The expected readings R_t in Equation (A10) is not equal to the actual total amount of reading in every period. The observed total amount of reading which is denoted by $\tilde{R}_t$ is defined by

$$\tilde{R}_t \ = \sum_{i=1}^{M} n_{it} r_{it}^* = \ \sum_{i=1}^{M} \sum_{j=1}^{J} n_{it} I\left(\zeta_i = \bar{\zeta}_j\right) \frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta}_j}{\alpha_2 / K_t + \kappa_{2j}} \nu_{it},$$

so it is obvious that

$$E\left(\tilde{R}_t\right) \ = E\left(E\left(\tilde{R}_t | n_{it}, \zeta_i\right)\right) = \ M \int_{s_{it}} \sum_{j=1}^{J} p_j p\left(n_{it} = 1 | s_{it}\right) \frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta}_j}{\alpha_2 / K_t + \kappa_{2j}} = R_t.$$

When the number M is large, we have $\tilde{R}_t/M$ is approximately equal to $E\left(n_{it} r_{it}^*\right) = R_t/M$ because of the law of large numbers. The expected average amount of reading per posting

$$y_t = \frac{R_t}{K_t} = \frac{R_t/M}{K_t/M} \approx \frac{\tilde{R}_t/M}{K_t/M} = \frac{\tilde{R}_t}{K_t},$$

which implies we can use the observed average amount of reading per posting to approximate the expected one in our model when the number of users is very large.

C Rational expectations

The following steps outline our approach to computing rational expectations and the resulting aggregate state transitions for the policy simulations and theoretical analysis.

1. Set structural parameters for utilities and costs of site usage, reading, and writing as well as μ^S, σ^S. Put bounds on state spaces of K_t^S, K_t^U, $\{k_{i,t}\}_{i=1}^N$, and y_t. This can be done by restricting value functions near lower and upper bounds of K_t^S, K_t^U, $\{k_{i,t}\}_{i=1}^N$, and y_t.

2. Guess the values for $\omega_0^U, \omega_1^U, \omega_2^U$ and $\omega_0^y, \omega_1^y, \omega_2^y$.

3. Discretize the state space and select points in the state space.

4. Solve for $p(n_{it} = 1 | s_{it}, \zeta_i)$, $p(r_{it} | s_{it}, n_{it} = 1)$, and $p(a_{it} | s_{it}, n_{it} = 1)$. The solution to dynamic choices require the value of y_t consistent with both aggregate reading and writing decisions (R_t and K_t). To get this value, we use the following steps:

(a) Choose an arbitrary y_t^{old} and $K_t^{U,old}$

(b) Solve for decisions by users: $\{n_{it}, r_{it}, a_{it}\}_{i=1}^{N}$. Equivalently, we compute $p(n_{it} = 1|s_{it}, \zeta_i)$, $p(r_{it}|s_{it}, n_{it} = 1)$, and $p(a_{it}|s_{it}, n_{it} = 1)$.

i. Given y_t^{old}, we can solve for $p(a_{it}|s_{it}, n_{it} = 1)$. If the state space is discrete, we use Rust (1987) to solve for EVs. If the state space is continuous or discrete but large, we use with Chebyshev approximation to expected value functions.

ii. Given $K_t^{U,old}$, we can solve for r_{it}^*.

iii. Given $p(a_{it}|s_{it}, n_{it} = 1)$ and r_{it}^*, we can solve for $p(n_{it} = 1|s_{it}, \zeta_i)$.

(c) Compute y_t^{new} and $K_t^{U,new}$. Check if $y_t^{old} = y_t^{new}$ and $K_t^{U,old} = K_t^{U,new}$. If the conditions hold then stop. If not, set $y_t^{old} = y_t^{new}$ and $K_t^{U,old} = K_t^{U,new}$ and iterate steps 4a-4c until convergence.

5. Solve for rational expectations by computing $K_{t+1}^U|K_t^U, K_{t+1}^S$ and $y_{t+1}|K_t^U, K_{t+1}^S$ and run OLS to get

$$\begin{aligned} K_{t+1}^U &= \tilde{\omega}_0^U + \tilde{\omega}_1^U K_t^U + \tilde{\omega}_2^U K_{t+1}^S \\ y_{t+1} &= \tilde{\omega}_0^y + \tilde{\omega}_1^y \left(K_{t+1}^U + K_{t+1}^S\right) \end{aligned}$$

6. Check if $\omega_0^U, \omega_1^U, \omega_2^U$ and $\omega_0^y, \omega_1^y, \omega_2^y$ are close to $\tilde{\omega}_0^U, \tilde{\omega}_1^U, \tilde{\omega}_2^U$ and $\tilde{\omega}_0^y, \tilde{\omega}_1^y, \tilde{\omega}_2^y$. If the conditions hold then stop. If not, replace ωs with $\tilde{\omega}s$ and iterate steps 2-5 until convergence.

Note that in estimation, the aggregate state transitions are observed and assumed to reflect the rational expectations in the current equilibrium, so no iteration to achieve the rational expectations is necessary. In policy simulations and theoretical analysis, however, we need to iterate to obtain them.

D Simulation Design

In section 4, we consider 2 segments of 3000 and 6000 users as their respective population sizes. We let both segments have the same cost of reading and heterogeneous costs ($\bar{\xi}_j$) of content generation. To simply the simulation, we assume there is no seasonal effect ($\bar{\tau}_{jt} = 0$

and $\bar{\kappa}_{1jt} = 0$). The reading cost parameters $\alpha_1 - \kappa_1 = 0.1$, $\alpha_2 = 1$ and $\kappa_{2i} = 0.0015$ imply a posting stock of $K_t = 10,000$ will induce an individual user to read 62.5 different postings per period. We let the cost of posting for Segment 1 be $\bar{\xi}_1 = 0.1$ and segment 2 be $\bar{\xi}_2 = 5$. Note that $\bar{\xi}_2$ is 50 times of $\bar{\xi}_1$, which implies Segment 2 has a much higher cost of posting and hence users in Segment 2 are likely to post much less than those in Segment 1. Indeed, we find in equilibrium a user in Segment 2 writes only about 2 postings in 100 periods whereas a user in Segment 1 writes about 350 posting in the same periods on average. We set the posting utility parameter $\gamma = 0.5$.

We endow every individual user with a randomly selected initial stock of user generated content. The initial aggregate stock of UGC is the summation of individual stocks plus a fixed initial stock of site generated content. The discount parameter β in the utility of posting is set to be 0.98. The site-generated content K_t^S is assumed to have an exogenous AR(1) process defined by equation(26) with $\omega_0^S = 1000$, $\omega_1^S = 0.5$ and $\sigma^S = 100$. Hence, K_t^S has a normal stationary distribution with mean equal to 2000 and standard deviation approximately 115. We use 2000 as the starting value for $K_{t=1}^S$.

We simulate individual postings and amount of reading for 100 periods. We then use the aggregate number of postings to re-estimate the dynamic law of motion for the posting stock, which will in turn lead to new values functions for both segments of user. The new value functions are used to simulate individual posting data again. This process is iterated until the law of motion for the posting converges. From numerous repeated experiments, we found it takes fewer than 20 iterations to converge to the rational expectations equilibrium. For illustration purpose, we show an example where the decay parameter ρ is set to be 0.6 (implying that each day there is a 60% chance that the post is noticed relative to the previous day).

E Model Estimation

E.1 Estimating the Reading Model

The individual-level reading model is

$$r_{it} = \frac{\alpha_1 - \kappa_{1t} - \zeta_i}{\alpha_2 / K_t + \kappa_2} \nu_{it}.$$

We assume that there are J segments and if user i is in the j-th segment, we have

$$r_{it} = \frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta}_j}{\alpha_2/K_t + \kappa_{2j}}\nu_{it} \tag{A12}$$

If we assume ν_{it} has the exponential distribution, the likelihood function for r_{it} given i in segment j is

$$Exponential\left(r_{it}|\frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta}_j}{\alpha_2/K_t + \kappa_{2j}}\right)$$

If we do not know segment membership of i, the likelihood becomes the following finite mixture distribution

$$\sum_{j=1}^{J} p_j Exponential\left(r_{it}|\frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta}_j}{\alpha_2/K_t + \kappa_{2j}}\right).$$

E.2 Estimating the Posting Model

One key component of estimation is to approximate the expected value functions in equation (16). This task is nontrivial for our model, because our state variables are mostly continuous with a wide support. Moreover, the control variable can take high-order discrete values. For this reason, we use Chebyshev approximation to approximate the expected value functions as described in (Dubé et al. (2009); Miranda and Fackler (2002)). Chebyshev approximation uses polynomial interpolation to approximate the expected value functions:

$$\tilde{EV}_j(s,a) \approx \psi\Gamma(s,a).$$

We can then rewrite the Bellman equation in the fixed point algorithm as a function of the interpolated functions

$$\psi\Gamma(s,a) = \int_{s'} \log\left(\sum_{a'\in A} \exp\left\{u(a'|s') - c(a'|s') + \beta\psi\Gamma\left(s',a'\right)\right\}\right) \cdot p(s'|s,a)ds'.$$

To compute the right-hand side of the above equation, we need to numerically evaluate an indefinite integral with respect to state transition probabilities of aggregate stock of posting. Since we use a normal distribution to model the probabilities, the Gauss-Hermite quadrature can be used to approximate the integration in the Bellman equation above (Miranda and Fackler, 2002). The Gauss-Hermite quadrature allows us to evaluate the integrand at fewer points than, for example, a Monte Carlo integration.

Once we compute both sides of the fixed point equation, we can formulate constraints to be used for our estimation based on the MPEC approach (Su and Judd, 2010):

$$R(s,a;\psi) = \psi\Gamma(s,a) - \int \log\left(\sum_{a'\in A} \exp\left\{u(a'|s') - c(a'|s') + \beta\psi\Gamma\left(s',a'\right)\right\}\right) \cdot p(s'|s,a)ds' = 0.$$

By approximating the expected value functions, we can transform a dynamic discrete choice model into a static one and use a maximum likelihood estimation to recover the structural parameters of our interest.

The joint likelihood of reading and posting for all individuals is then

$$\left\{\prod_{i=1}^{M}\sum_{j=1}^{J} p_j \prod_{t=1}^{T} Exponential\left(r_{it}|\frac{\alpha_1 - \kappa_{1jt} - \bar{\zeta_j}}{\alpha_2/K_t + \kappa_{2j}}\right) \frac{u(a_{it}|s_{it})}{\sum_{a'\in A}\exp\left\{u(a'|s_{it}) - c(a'|s_{it}) + \tilde{EV}_j(s_{it},a')\right\}}\right\}. \tag{A13}$$

The direct MLE approach (e.g., Kamakura and Russell (1989)) is applied to estimate the parameters. To compute the standard errors of parameter estimates in the posting model, we use nonparametric bootstrapping. Note that we allow for heterogeneity for reading and posting costs using finite mixture models, which makes it difficult to implement nonparametric bootstrapping for computing standard errors due to the label switching problem. Geweke and Keane (1997) propose labeling restrictions that prevent the components of the mixture from interchanging across bootstrapped samples. For example, segments can be ordered according to their sizes to preserve segment labels consistently across bootstrapped samples.

E.3 Estimating the Site Visitation Model

Lastly, we have the likelihood function for site-visitation data following equation (25):

$$\left\{\prod_{i=1}^{M}\sum_{j=1}^{J} p_j \prod_{t=1}^{T}\left[P(n_{it}=1|s_{it})^{n_{it}} P(n_{it}=0|s_{it})^{1-n_{it}}\right]\right\}, \tag{A14}$$

which is also estimated by MLE.